The role of the 3D architecture of the nucleus in shaping vertebrate transcriptional regulation

El papel de la estructura 3D de la cromatina en la evolución de la regulación transcripcional de vertebrados

Rafael Domínguez Acemel

Director: José Luis Gómez-Skarmeta

Doctorado en Biotecnología, Ingeniería y Tecnología Química
Universidad Pablo de Olavide

Regulación génica y morfogénesis
Centro Andaluz de Biología del Desarrollo (CSIC/UPO)
Sevilla

Portada y Contraportada de Antonio González García (Instagram @watercolorbyantonio).

A mis padres y abuelos.

Porque ignoraba que el deseo es una pregunta/ cuya respuesta no existe
una hoja cuya rama no existe/ un mundo cuyo cielo no existe.

Luis Cernuda.

May your hands always be busy/ may your feet always be swift
may you have a strong foundation/ when the winds of changes shift.

Bob Dylan.

Avanti con la guaracha.

Silvio Fernández.

Abstract/Resumen

Animal morphological diversity is astonishing and it is partially due to differences in gene expression between different species during development. Recently, the genome folding in Topologically Associated Domains (TADs) found in most animals has been shown to be critical in the control of transcription during development. Distal enhancers are able to interact with the promoters of developmental genes only when they belong to the same 3D environment or TAD. Therefore, we investigated how changes in the 3D folding of the genome could have impacted the changes in gene regulation responsible for the evolution of vertebrates. In order to do so we combined syntenic analysis with Chromatin Conformation Capture experiments such as 4C-seq and HiChIP.

First we compared the chromatin folding around the zebrafish HoxD and the amphioxus Hox loci using 4C-seq experiments coupled to computational modelling. The chromatin architecture around the vertebrate HoxD locus is peculiar, with all the HoxD genes located at the boundary between two TADs allowing them to switch to respond to distal enhancers located in either of the two TADs during the patterning of the limbs. In contrast, all the amphioxus Hox genes belong to the same TAD. However, the region located downstream from *Hox1* is homologous to the vertebrate anterior TAD and is wired both in 3D and functionally to the regulation of Hox genes also in amphioxus. This suggests a stepwise evolution of the chromatin folding in two TADs found in extant vertebrates, with the anterior TAD being already wired to Hox genes in the last common ancestor of chordates.

Second we performed a genome wide comparison of the chromatin folding between zebrafish and amphioxus using HiChIP and antibodies against different histone modifications. Using H3K4me3 HiChIP experiments we were able to identify the Regulatory Landscapes (RLs) of all active developmental promoters using a single experiment. By doing so we were able to identify almost four hundred cases of chromosomal rearrangements that potentially altered the boundaries of a TAD and were susceptible to generate regulatory novelties in the vertebrate lineage. Also, we found that the two events of whole genome duplication that occurred at the root of vertebrates allowed some of the paralog genes originated to increase their RLs both in size and in number of enhancers.

Keywords: TADs, regulatory landscapes, 4C-seq, HiChIP, origin of vertebrates, amphioxus, whole genome duplications, Hox genes

La diversidad morfológica que encontramos en el reino animal es impresionante. En parte, esta diversidad morfológica responde a diferencias de expresión génica entre diferentes especies durante el desarrollo embrionario. Recientemente se ha demostrado que la organización tridimensional del genoma en dominios de asociación topológica o *TADs*, presente en la mayoría de animales, es esencial en el control de la transcripción durante el desarrollo. Los potenciadores de la transcripción o *enhancers* son solo capaces de activar la transcripción de genes situados dentro de su mismo *TAD*. Por ello, nosotros hipotetizamos que cambios en la estructura de los *TADs* pudieron conllevar cambios en regulación génica importantes en el origen de los vertebrados. Para ello combinamos análisis de sintenia con experimentos de Captura de la Conformación Cromosómica (*Chromosome Conformation Capture*) como *4C-seq* o *HiChIP*.

Primero comparamos la arquitectura de la cromatina alrededor de los genes HoxD de pez cebra y los genes Hox de anfioxo utilizando experimentos de *4C-seq* y modelos por ordenador. La estructura de la cromatina alrededor de los genes HoxD en vertebrados es peculiar, puesto que estos genes se encuentran en el borde entre dos *TADs*. De esta forma son capaces de responder a *enhancers* localizados tanto en el *TAD* anterior como en el posterior, lo cual es crítico durante el desarrollo de las extremidades de vertebrados (ya sean patas o aletas). Sin embargo, en anfioxo, todos los genes Hox se encuentran situados dentro de un mismo *TAD*. Es interesante comprobar aun así que la región que encontramos aguas abajo del gene *Hox1* de anfioxo es homóloga al *TAD* anterior de vertebrados y que esta región está también conectada con la regulación de los genes Hox en amphioxus tanto a nivel estructural como funcional. Todo ello sugiere que la arquitectura en dos *TADs* que encontramos en vertebrados es el resultado de una evolución por pasos en la que primero se asoció el *TAD* anterior en el ancestro de cordados para luego asociarse el posterior.

Por otra parte realizamos también un análisis global de la evolución de la arquitectura del genoma entre pez cebra y anfioxo usando *HiChIP* y anticuerpos dirigidos contra modificaciones de las colas de las histonas. Usando experimentos de *HiChIP* contra la modificación H3K4me3 identificamos los paisajes reguladores de todos los genes activos durante el desarrollo tanto en pez cebra como en anfioxo. Con ello fuimos capaces de encontrar casi cuatrocientos casos de rearreglos cromosómicos que pudieron conllevar la rotura de un borde de *TAD* y por tanto pudieron generar novedades a nivel regulatorio en el linaje de los vertebrados. Además, descubrimos que los eventos de duplicación de genoma completo que ocurrieron en la base de vertebrados permitieron a algunos genes paralogos incrementar sus paisajes reguladores tanto en tamaño como en número de *enhancers*.

Palabras clave: *TADs*, paisaje regulador, *4C-seq*, *HiChIP*, origen de vertebrados, anfioxo, duplicación de genoma completo, genes Hox

Contents

Chapter 1

Introduction

1.1

A genetic explanation to the evolution of animal diversity

At the beginning of 2018 more than 1,200,000 extant species of animals were indexed in the Catalogue of Life (Roskov et al. 2018), of which less than 70,000 (5.7%) were vertebrates. Beyond these figures, animal diversity and its beauty is evident and has drawn the fascination of curious people for centuries. Their morphological diversity has allowed animals to behave in many different ways and to thrive in the most disparate environments of Earth. However, and despite all these differences, many clues have always hinted that we (animals, metazoans) belonged together in some way. This picture became more and more evident after the development of cell and molecular biology fields.

In 1859, and without this information at hand, Charles Darwin was one of the first in successfully propose the idea that all animals share a common ancestor and that morphological diversity and speciation is due to inheritable random variations between individuals that are then selected (Figure 1.1A). This natural selection occurs according to the fitness of the new characteristic within a given environment (Darwin 1861). His paradigmatic example was the relationship in the shape of the beaks of the Darwin finches with the food sources available in the different Galápagos islands (Figure 1.1B). One of the first challenges then was explaining the origin of such morphological variability and how these variations could be inherited. New advances such as Gregor Mendel's laws of character inheritance (Mendel 1946), the discovery of the DNA as the basic molecule carrying inheritable information (Avery, MacLeod, and McCarty 1944) and the work of Thomas Morgan and others linking the mendelian inheritance of several traits in *Drosophila* to changes in certain regions of the *Drosophila* DNA (Morgan 1916, Figure 1.1C) helped to clarify the picture. Since mutations in the DNA were considered to be random, they could be inherited and they also determined external traits, everything fitted perfectly into the Darwinian model.

Another critical step was to explain the connection between the genotype (the inheritable DNA molecules, the genome) and the phenotypes we observe (i.e. the organism morphology). For that it was key to understand how the genome holds and use the information on how to synthesize proteins. Each of the segments of the genome encoding the synthesis of a given protein is called a protein coding gene. During the process of transcription these genes are transcribed to a messenger RNA

(mRNA) intermediary that is exported to the cytoplasm where they instruct the ribosomes in the synthesis of proteins in a process called translation (Brenner, Jacob, and Meselson 1961; Watson 1963). Proteins are the building blocks of many cellular components and the effectors of many of the physiological processes happening in the cells. Therefore, the protein content (proteome) of a given cell very much influence its shape, behavior and function.

Whereas DNA content and sequence is roughly equal between the different cells of a multicellular organism and even between cells of different organisms belonging to the same species, the mRNA content (also called transcriptome) and subsequently the proteome of cells may vary drastically from one cell to another. There are many kinds of stimuli able to change the proteome of a cell including presence or absence of nutrients (Jacob and Monod 1961), changes in temperature (Grossman, Erickson, and Gross 1984) and cell to cell communication events (Von Ohlen et al. 1997). There is also a number of mechanisms that control the protein composition of cells at different levels, or said in an almost equivalent way, the gene expression. Among them, regulation at the level of transcription (i.e. deciding which genes of the genome are going to be transcribed to mRNA and how many mRNA copies of each of them) is of paramount importance and one of the best studied examples. Recent advances that allow to examine mRNA composition with single cell resolution have shown that the transcriptome of a given cell is sufficient to infer its identity and function (Farrell et al. 2018; Sebé-Pedrós et al. 2018; Rosenberg et al. 2018). It is most likely that the transcriptomes of these cells are not just a mere characteristics that allow to classify them, but play a crucial role in determining cell identity and future fate. In addition, it has been quantified that the transcriptome of mouse fibroblasts explain up to 40% of their final proteome (Schwanhüusser et al. 2011). However, we should bear in mind that there are many other regulatory layers controlling the final proteome of animal cells beyond transcription, such as the control of mRNA stability and degradation (Giraldez 2006), the choice between protein isoforms thanks to alternative splicing (Barbosa-Morais et al. 2012), the regulation of the mRNA translation efficiency (Xue et al. 2015), the postranslational modification of proteins (Schroeter, Kisslinger, and Kopan 1998) and the control of the degradation rate and the subcellular localization of these proteins (Elosegui-Artola et al. 2017). All of these processes (among others) are conserved across animals and are tightly regulated.

Since we are interested in the evolution of animal morphology we are mostly concerned about concerted changes in proteomes taking place during animal development, which is the critical step bridging the inheritance of a given genome with the elaboration of an adult organism with fixed morphological characteristics. The resultant multicellular organism is composed by many cells of different kinds, precisely located, and specialized in performing disparate functions and in producing different sets of proteins from the original DNA sequence they share. To reach that point, the original zygote and its progeny took a series of decisions including the possibility of dividing (Bessa et al. 2002), migrating (Sánchez-Higueras and Hombría 2016), changing their shape (Nicolás-Pérez et al. 2016), communicating to neighboring cells (Dominguez-Cejudo and Casares 2015) or even dying in a controlled manner (Pajni-Underwood et al. 2007). These decisions were instructed to a great extent by the proteins they produced through the developmental process, and changes in those decisions may end up in the evolution of different morphologies. Then, theoretically, variation in developmental processes leading to morphological changes can happen both through the modification, appearance or disappearance of proteins acting during development or through the modification of regulatory mechanisms that control gene expression (Figure 1.1D). It is worth noting that both the information about proteins and about how to deploy the regulatory

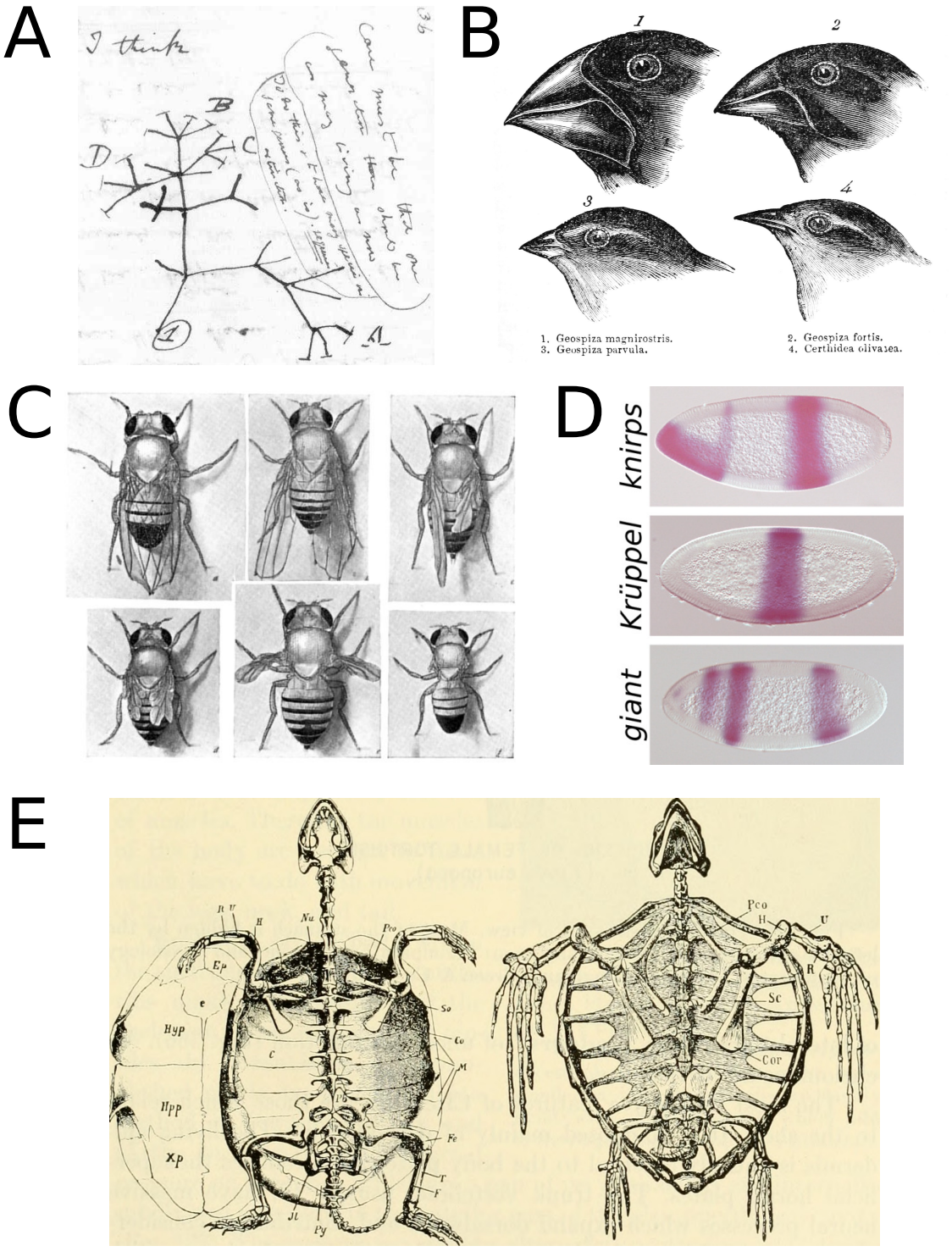


Figure 1.1: The astonishing morphological diversity and its origin. In (A) is one of the first known sketches of a tree of life, from Charles Darwin's *First Notebook on Transmutation of Species* (1837). In (B) there are drawings from several Darwin finches heads, with beaks of species 1 and 2 being adapted to crack nuts and seeds while species 3 and 4 eat mostly insects. Drawing from Darwin's *Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle [...]* (1845). In (C) there are several wing mutants described and mapped to different chromosomic positions: cut, beaded, stumpy, stumpy, vestigial and apterous. From Thomas Hunt Morgan *A Critique of the Theory of Evolution* (1916). In (D) the result of the in-situ hybridisation of three *Drosophila* GAP genes in early embryos (the pink color reveal the location of the mRNA of each of the genes). These genes are expressed in specific segments that do not develop if they are absent, highlighthing the importance of gene regulation in development. In (E) there is the drawing of both a terrestrial and a marine turtle skeletons, from Herbert Rand's book *The Chordates* (1950).

mechanisms listed before are in one way or another written in the DNA sequence, as we will explore further in following sections. Therefore, in principle all these processes are susceptible to evolve following Darwin's principles.

Indeed, many of the elements of Darwin's theory are widely accepted today such as the common ancestry of all living beings and the unquestionable importance of natural selection. However, there are certain assumptions that have been challenged and one of them is precisely related to the cause and process by which morphological novelties appear. According to Darwin, morphological novelties appeared slowly with many intermediate states that needed to be also favored by natural selection, even though he found it difficult to find adaptive value to those intermediates states. Notwithstanding, there are many examples of seemingly sudden novelties such as the turtle shell (Wang et al. 2013, Figure 1.1E) or the insect wings (Clark-Hachtel and Tomoyasu 2016) that do not seem to accommodate with the timing of this hypothesis. Another yet more incredible example is the rapid emergence of the body plans of all extant bilaterian phyla during the Cambrian explosion, body plans that have remained largely conserved ever since (Davidson and Erwin 2006). In order to have a better understanding on how those events took place we will study the transcriptional regulation during the development of chordates in a comparative way. Then we will try to infer how changes in gene expression may have contributed to the evolution of the vertebrate body plan which comprise an important number of morphological novelties (Holland, Holland, and Holland 2015).

1.2

Transcriptional regulation in animals

The central question in transcriptional regulation is understanding how the RNA Polymerase II (RNAP II) is specifically recruited to the beginning of the genes (transcriptional start sites, TSSs) to produce the adequate amount of mRNA molecules from each of these genes. In the context of development we wonder how different cells in the embryo transcribe different sets of genes. In order to solve those questions we need to explore different factors involved. These factors include the RNAP II complex itself (Brookes and Pombo 2009), many specific DNA sequences (cis-regulatory elements, CREs), many proteins such as histones, transcription factors and chromatin remodellers (accounting for about the 10% of all the proteins encoded in mammalian genomes; Fulton et al. 2009), the epigenetic modification of DNA and histones and physical properties such as the 3D-folding of the chromosomes.

1.2.1

Cis-regulatory elements: core promoters and enhancers

We will focus first on CREs, which comprise several entities such as core promoters, enhancers and silencers. Core promoters are the regions that surround the TSSs of genes, typically extending 40bp both upstream and downstream. Indeed, their function is to direct the RNAP II to their TSS and establish the proper directionality of transcription (Haberle and Lenhard 2016). In that regard, the development of CAGE-seq experiments was really important in order to understand TSS placement

within promoters because it allows to pinpoint the exact position of TSSs by carefully sequencing just the mRNA portion adjacent to the 5' cap, a modified nucleotide that protects the anterior end of a mRNA molecule (Shiraki et al. 2003). By applying this technique to several mammals (Carninci et al. 2006) and *Drosophila* (Hoskins et al. 2011) it has been demonstrated that most animal promoters display a broad range of possible TSSs within a 20-30bp window (being classified as broad promoters) while just a minority of them display the expected pattern of one TSS per promoter (classified as sharp). These TSS patterns are relevant for transcriptional regulation as we will explore later. Regardless of its type, every core promoter must be able to direct the onset of transcription by recruiting the transcriptional pre-initiation complex (PIC) composed of the RNAP II and a group of general transcription factors (i.e. proteins regulating transcription by directly binding to DNA) like the TFIID complex (Haberle and Lenhard 2016). This TFIID factor recognises some of the sequence signatures that characterize core promoters. Among them it is important to highlight the Inr, which is the most common promoter signature found both in mammals and in *Drosophila*. Its common motif consistS in a di-nucleotide of a pyrimidine (C or T) followed by a purine (A or G) that overlaps precisely with the TSS. It can be found alone or in combination with other elements such as the TATA-box or the DPE. The TATA-box, despite being the most classic example of promoter element, is only present in 15% of mammalian promoters, often belonging to the sharp category. In those cases, the TATA motif is found precisely 28 to 32 bp upstream from the TSS, playing a role in stabilizing the interaction of the promoter with the TFIID transcription factor. On the contrary, the DPE that is commonly found around *Drosophila* promoters (and to a lesser degree also in mammals) seems to stabilize the TFIID from 28 to 32 bp downstream from the TSS. Interestingly, TATA and DPE elements are not commonly found together within the same promoters (Burke and Kadonaga 1996) although they are commonly associated with the sharp category (Hoskins et al. 2011). Finally, another common feature at the level of sequence is the accumulation of CG di-nucleotides around vertebrate promoters (known as CpG islands), feature that is commonly associated to the big group of broad promoters (Carninci et al. 2006, Akalin et al. 2009).

Therefore, core promoters are crucial for gene regulation by directing the RNAP II to the TSS, but so far we have not presented any element that explains how different cells activate and repress the expression of different sets of genes. This is achieved because some core promoters are able to integrate the regulatory information coming from enhancers, another type of CREs. Enhancers are sequences that contain binding sites for a combination of transcription factors (TFs) that are expressed in a tissue/temporal specific manner in contrast to general transcription factors such as TFIID (reviewed in Visel, Rubin, and Pennacchio 2009). Only in those cells where the correct combination of TFs is present and bound to the enhancer transcription may be activated from the TSS of target core promoters. This system allows enhancers to accurately drive the expression of particular genes to specific cell populations within the embryo. An important characteristic of the tissue specific regulation through enhancers is its modularity, since the same gene can be activated by different enhancers in different cell populations (see Figure 1.2A). Developmental genes, that often need to be precisely transcribed in different tissues and developmental stages, usually display the largest amount of enhancers (Calle-Mustienes et al. 2005). In contrast to promoters, enhancers can appear in much wider genomic territories both upstream and downstream from the target TSS, and this territory is commonly referred to as the regulatory landscape (RL) of a given gene. The RL of some developmental genes in mammals can expand more than a megabase (e.g. *HoxD* genes, *Shh* and *Irx3/Irx5*; Montavon et al. 2011, Symmons et al. 2016, Smemo et al. 2014) and this raises

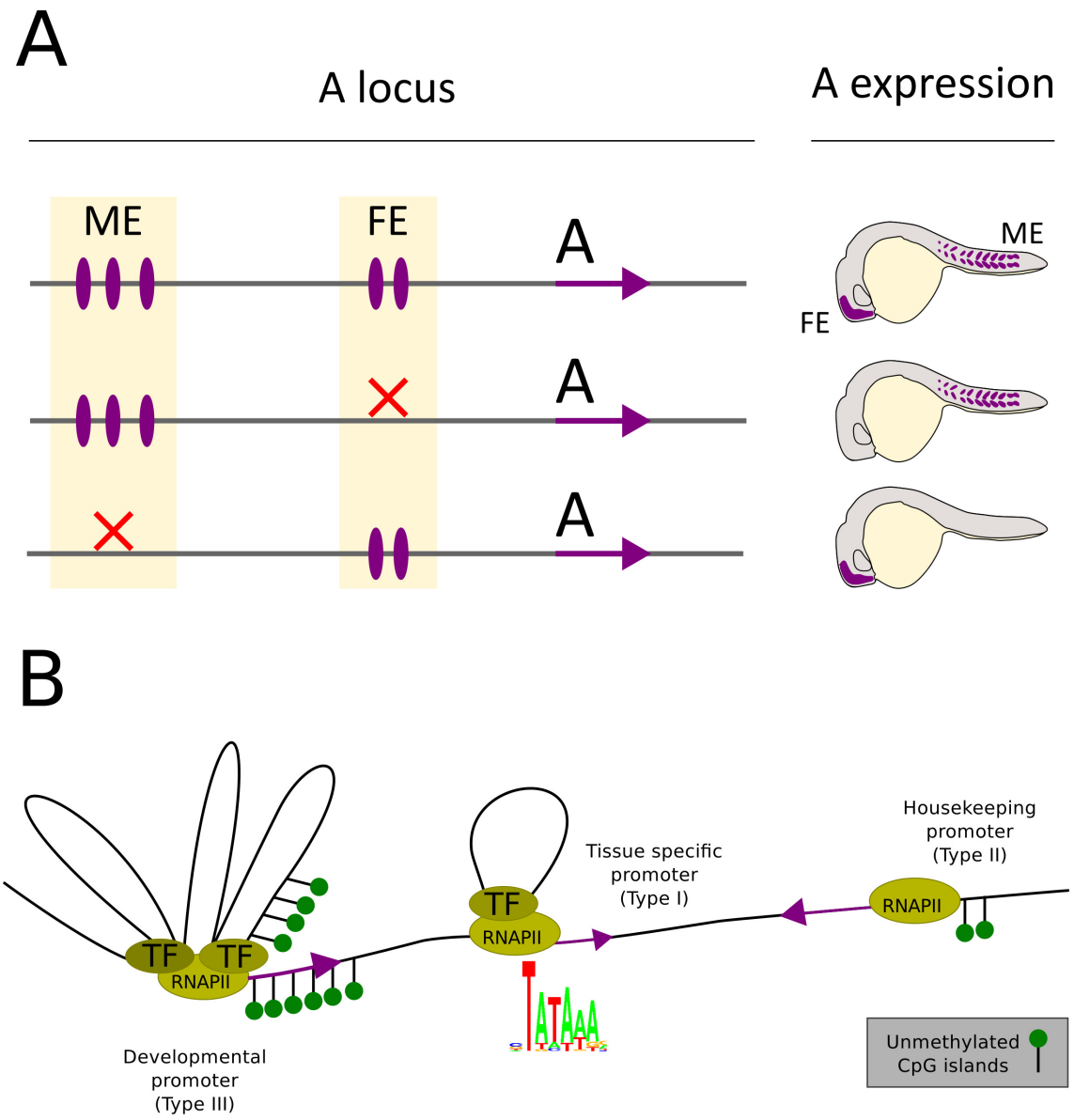


Figure 1.2: Interplay between enhancers and promoters. In (A) the cartoon shows the classic model of enhancer modularity. The complex pattern of expression of a given developmental gene (here termed gene A, expressed in forebrain and in the somites) is the result of the combinatory effect of forebrain enhancers (FE) and muscle enhancers (ME). In (B) some hallmarks of three categories of animal promoters are represented, inspired in the work by Haberle and Lenhard 2016. Housekeeping promoters do not tend to integrate regulatory information from enhancers, display a broad TSS distribution and tend to have unmethylated CpG islads in the TSS surrounding area. Tissue specific promoters usually display a sharp distribution of the TSS, likely due to the presence of a well positioned TATA-box motif, and respond to a limited set of tissue specific enhancers. Finally, developmental promoters: (i) integrate the information from many enhancers, (ii) display an even broader TSS distribution when compared with housekeeping promoters and (iii) contain large CpG island regions that extend towards the gene body.

several questions such as how TFs bound to an enhancer located hundreds of kilobases away from the core promoter are able to trigger transcription. Seminal studies done in the murine beta globin locus showed that beta globin enhancers located inside the LCR come close to the beta globin promoters they regulate in the 3D space, looping out the intervening DNA sequence (Tolhuis et al. 2002). TFs like Gata1 and Ldb mediate this looping that its indispensable for the transcription of the beta globin genes in mature mouse erythrocites (Palstra et al. 2003). However it is still a matter of debate if a stable tethering or just proximity is needed for enhancers and promoters to communicate productively (Deng et al. 2012). The fact that many enhancers reside inside the introns of the same genes they regulate suggests that the binding might be transient and dynamic (Spieler et al. 2014). In any case, enhancers and their target promoters tend to cluster together in space despite they can be located far apart in the linear sequence, and therefore the 3D folding of DNA plays a crucial role in the regulation of gene expression mediated by enhancers. For that reason, the following section will be dedicated to nuclear architecture exclusively.

Another reasonable question that arise from the fact that enhancers can be placed far apart from their target promoters is how specificity is achieved, since in a 1 Mb window we often find some other core promoters much closer than the target promoter. However, their regulation is often totally independent from the activity of the enhancers. It is true that the big RLs of developmental genes, at least in vertebrates, tend to be gene poor and comprise big chunks of what was called gene deserts (Ovcharenko et al. 2005). In addition, the folding of the DNA can favor some enhancer-promoter contacts above others as we will elaborate later. Nevertheless, there are some cases where those circumstances do not seem to fully explain the specificity observed, for instance in the rather extreme case of the *Fgf8* regulation. *Fgf8* is expressed in a myriad of very specific cell populations during development and for that its promoter receives the input of many enhancers located inside a big RL, spanning 600kb in mouse (Marinić et al. 2013). This RL contains seven other genes that do not seem to respond at all to the activity of *Fgf8* enhancers, and some of them are just broadly and evenly expressed throughout all the embryo. An interesting study using STARR-seq in *Drosophila* has shed some light on this topic of enhancer-promoter specificity (Zabidi et al. 2014). Briefly, these experiments allow to examine which fragments of the genome are able to function as enhancers of a given promoter in a given cellular context. The first step is to generate a library of constructs containing a promoter of interest followed by random fragments of DNA covering the entire genome. This library is then introduced in a given cell line, and those constructs containing enhancer sequences that are compatible with the promoter and the TFs present in the cell line are able to trigger their own transcription. The RNA of these cells, containing the sequences of active enhancers, is then sequenced and the enhancer activity of each DNA fragment is quantified. Following this procedure using seven different promoters and two different cell lines it was shown that the seven promoters could be classified in two groups: developmental and housekeeping. Developmental promoters were able to read information coming from enhancers that: (i) only showed activity in one of the two cell lines, (ii) contained binding sites from cell type specific TFs, (iii) could be several kb apart from the gene they regulate in their endogenous context. In general, those are the classic enhancer features we described earlier. In stark contrast, housekeeping enhancers usually showed activity in both cell types, contained binding sites for broadly expressed TFs and were located just adjacent to the core promoters of the genes they regulate in their endogenous loci. Since both kind of promoters were able to respond to a great number of enhancers when they were exposed to them, it seems unlikely that there is a high degree of specificity between enhancers and promoters. Rather, this seems to favor more an scenario in

which the core promoters of developmental genes tend to be the only ones inside their RLs able to integrate regulatory information coming from distal enhancers.

There are no equivalent STARR-seq experiments performed in vertebrates, but it is also possible to classify the core promoters of genes that are broadly expressed and the ones that are tightly regulated by distal enhancers in at least three different groups based on some of the core promoter features described earlier (reviewed in Haberle and Lenhard 2016). Type I promoters are commonly found in genes that are expressed in specific differentiated adult cells. They are the ones displaying a well positioned TATA-box motif and a sharp TSS distribution and are able to integrate signals from few cell type specific TFs through enhancers that are close to the TSS. Type II promoters are those of housekeeping genes and they present broad TSS distributions, no TATA-boxes and in the case of vertebrates CpG islands surrounding these TSS distributions. They do not usually integrate context information of cell type specific TFs since they regulate genes that ought to be transcribed everywhere. Finally, type III promoters are those of developmentally regulated genes and are strikingly more similar to type II promoters than they are to type I. They present even broader distributions of possible TSSs, no TATA-boxes and CpG islands that expands from the promoter well inside the gene body of the developmental gene. However, in stark contrast with type II promoters, they are able to integrate multiple context specific inputs of many distal enhancers (see Figure 1.2B). It is important to note that regulation through distal enhancers appears to be a critical novelty of animals since it is not present in other sister holozoans such as *Capsaspora owczarzaki* (that only have type II promoters, Sebé-Pedrós et al. 2016), but likely present at least in the last common ancestor of bilaterians and cnidarians since enhancers are easily found in the sea anemone *Nematostella vectensis* (Schwaiger et al. 2014).

1.2.2 Epigenetic control of CRE accessibility

Another important factor involved in animal transcriptional regulation is related to the accessibility of both the RNAP II and the TFs to the DNA and how it is modulated by different epigenetic factors. For instance, one of the mechanisms controlling the binding of proteins to DNA and transcription in general consists in the direct methylation of the cytosines of the CpG di-nucleotides (reviewed in Bogdanović and Lister 2017). Most vertebrate CG di-nucleotides are methylated by the DNMT protein family of methyltransferases, except for those that accumulate in CpG islands around the broad type promoters (i.e. types II and III) and enhancers. The TET proteins are the ones in charge of selectively demethylate CpG islands and enhancers. In general, methylation of enhancers and promoters results in a strong repression of the transcription of their target genes. Moreover, the tight regulation of both the methylation and demethylation of enhancers and promoters plays a crucial role in several developmental processes such as the control of gastrulation (Bogdanović et al. 2016) and the specification of the germ line cells in vertebrates (Hargan-Calvopina et al. 2016). Outside vertebrates, cytosine methylation occurs mainly along the gene body of actively transcribed genes and in transposons, both features being likely ancestral since they are conserved along eukaryotes from animals to plants (Feng et al. 2010, Wang et al. 2014, Marlétaz et al. 2018).

Additional layers controlling the access of TFs to CREs rely on the fact that the eukaryotic DNA is tightly associated around proteins called histones forming nucleosomes. Each core nucleosome

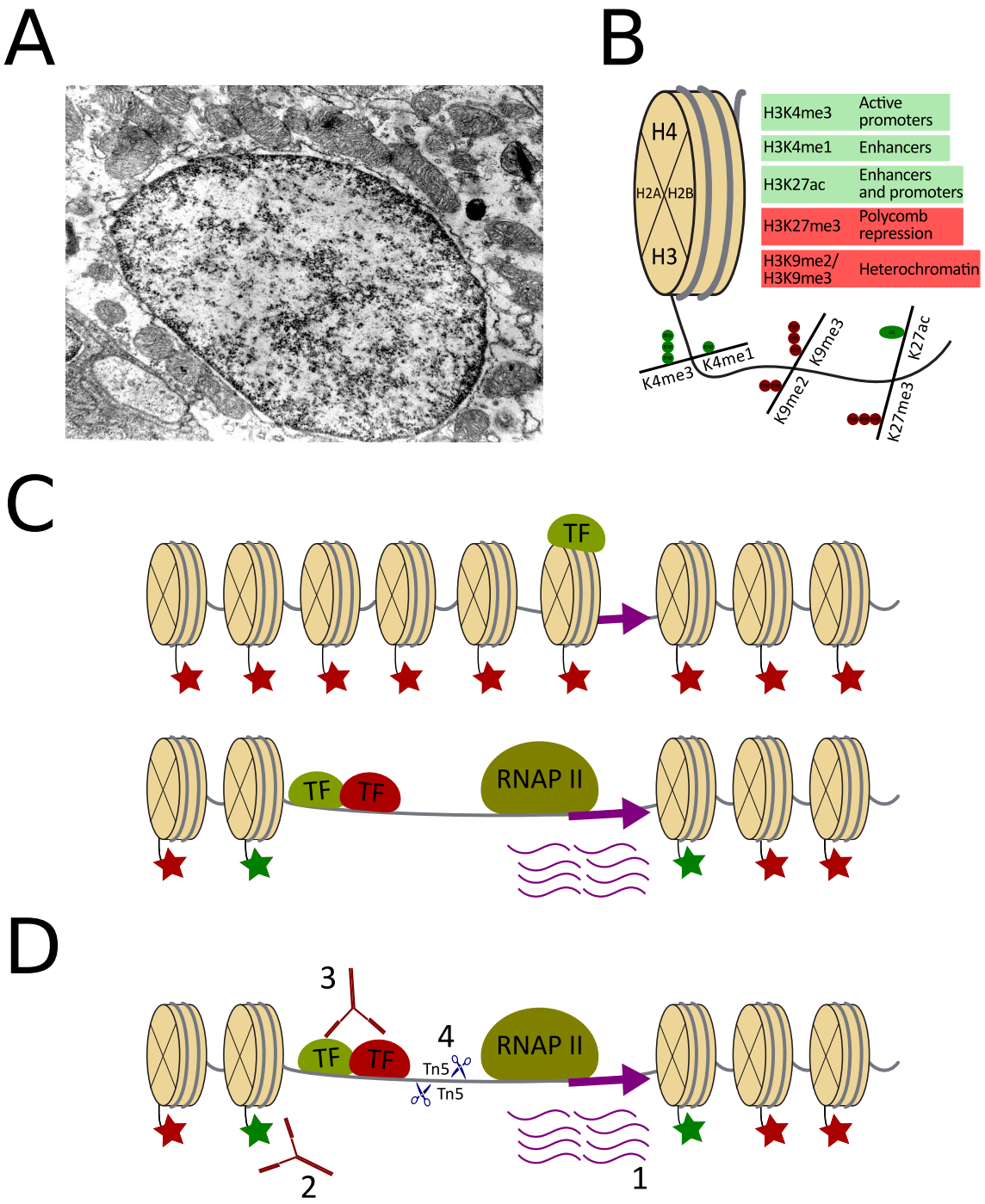


Figure 1.3: The accessibility to CREs is critical for their function. (A) EMT microscopy image of the nucleus of a chloride cell (present in the gills of several fish species) where dark spots of heterochromatin both near the nuclear envelope and the nucleolus are readily visible (photography by T.Voekler, CC BY-SA 3.0 license). (B) Scheme showing a nucleosome composed by the DNA helix wrapped around the histone octamer. CREs wrapped around a nucleosome are hardly accessible by TFs and will hardly exert a function. The histone 3 tail modifications covered in the text are also depicted with a summary of their regulatory role. Activating and repressive modifications are colored green and red respectively. (C) Cartoon showing how a pioneer TF (in green) is able to bind to a nucleosome associated DNA, recruit chromatin regulators that modify the histones, and expose the DNA in order to activate a given target gene. (D) Several examples showing how NGS based techniques allow to study gene regulation at different levels: (1) RNA-seq allows to sequence the resulting RNA molecules, (2-3) ChIP-seq experiments use antibodies to target histone modifications or tissue specific TFs respectively and ATAC-seq (4) takes advantage of the fact that the Tn5 transposases cut more often in open chromatin regions such as those surrounding active CREs.

is composed of four different histone dimers (H2A, H2B, H3, H4) and the 146 bp of DNA that is packed around them (Richmond et al. 1997). The regions of the DNA that are in between two nucleosomes are called linker DNA and they span around 80 bp. In general, if a CRE is associated to a nucleosome core it will not be accessed easily by most TFs and therefore it will not be able to have a role in transcription. The nucleosome remodelling processes that are necessary to expose or hide enhancers and promoters to the TFs are tightly regulated and many of them are also largely conserved across metazoans and even beyond (Schones et al. 2008, Lee et al. 2007). Here we will focus on a number of post translational modifications happening in a rather disorganized part of the histone proteins that are called histone tails. Epigenetic modifications in these tails play an important role in the interaction of histones with the DNA and with different protein complexes, and therefore in gene expression (Strahl and Allis 2000). Particularly we will concentrate on some modifications happening in the lysine residues located in the positions 4, 9 and 27 of the H3 (H3K4,H3K9 and H3K27 respectively: see Figure 1.3B).

First of all we will explore epigenetic modifications leading to the repression of transcription such as the H3K9 dimethylation and trimethylation signature (H3K9me2 and H3K9me3 respectively). These marks occupy large portions of the genome and are associated with a high level of compaction of the nucleosomes, the methylation of CpG islands and the absence of transcription (reviewed in Becker, Nicetto, and Zaret 2016). The big size and the characteristic level of compaction of these regions facilitated their description using microscopy almost a century ago (see Figure 1.3A), and they were termed heterochromatin (chromatin being the physiological association of the chromosomes with nuclear proteins). A large fraction of the genomes is always associated to this repressive mark regardless of the cell type or cell lineage (constitutive heterochromatin), often comprised by repetitive rich regions including telomeres, centromeres and transposable elements. However, the *SETDB1* protein involved in the H3K9 methylation can also be specifically recruited by cell type specific TFs (e.g. the KRAB family of TFs in vertebrates, Liu et al. 2014) to silence specific genes. The role of SETDB and the H3K9me3 mark in transcriptional silencing is conserved at least in animals and fungi.

Another epigenetic mark involved in the silencing of gene expression is the trimethylation of the lysine 27 (H3K27me3) that is catalyzed by the EZH family of proteins of the Polycomb Repressive Complex 2 (PRC2, reviewed in Schuettengruber et al. 2017). Both this epigenetic mark and the PRC2 are conserved across eukaryotes including plants and fungi. However, the recruitment of the PRC2 complex to the loci that need to be silenced differs between *Drosophila* and mammals, for example. In *Drosophila*, specific sequence motifs known as PREs (Polycomb Response Elements) bind several TFs that in turn recruit both the PRC1 and the PRC2 complexes. In mammals, however, the main target of PRC complexes are hypomethylated CpG islands, specially if they are devoid of activating TFs. The recruitment to these islands is at least partially mediated by the KDM2B CxxC domain that recognize the unmethylated CG pattern and has also affinity for the ncPRC1 (non canonical PRC1). This ncPRC1 is able to ubiquitinate the H2A tails (H2AK119ub) and this modification is recognized by the JARID2 TF that in turn recruit the PRC2. Then the PRC2 member EZH is able to deposit the H3K27me3 modification that is further recognized by the CBX factor of the PRC1 complexes closing a stabilizing feed forward loop. The presence of both polycomb complexes in turn difficult transcription through the compaction of the nucleosomes mediated by some of their members like the PRC1 PHC proteins that are able to polymerize thank to their SAM domains. Interestingly, these polymerization events seem to generate long range contacts between distant polycomb repressed genes at least in mammals (Vieux-Rochas et al.

2015). It is also important to note that the H3K27me3 modification take place at the same lysine residue that the H3K27ac modification, which is characteristic of active promoters and enhancers as we will describe later.

Next we will explore activating epigenetic modifications of nucleosomes starting from the monomethylation and the trimethylation of the lysine 4 (H3K4me1 and H3K4me3 respectively). The H3K4me3 mark is strongly enriched in the two nucleosomes that are located immediately before and immediately after the nucleosome depleted region characteristic of active promoters, and this pattern is conserved across all eukaryotes (reviewed also in Schuettengruber et al. 2017). The precise location of these two nucleosomes is more evident in those promoters categorized as broad (types II and III, Haberle and Lenhard 2016). H3K4me3 modified histones have been shown to interact with proteins associated with chromatin remodeller complexes such as CHD1 and BPTF, that are able to reposition nucleosome cores and expose the DNA both to TFs and to the PIC. Moreover, it is able to recruit H3K27 acetylases such as CBP that promotes the H3K27ac modification, that is characteristic of active genes and counteracts the repression mediated by polycomb. Interestingly, it has been shown that some promoter associated nucleosomes are marked both with the activating H3K4me3 and the repressive H3K27me3 epigenetic marks in several vertebrates (reviewed in Voigt, Tee, and Reinberg 2013). Those bivalent promoters are often related with developmental genes that are in a poised state, being able to undergo both repression on activation upon the correct developmental signals. It has been proposed that this poised state allows a sharp and strong transcriptional activation when activating signals reach a given threshold, in contrast to an scenario where the activation is gradual. The first pattern might be more useful for the robust regulation of some developmental processes. The methylation of the H3K4 is catalized by several COMPASS protein complexes (one of the two main types of complexes of the Trithorax group of proteins, the other is the SWI/SNF), that are partially shared by all eukaryotes (Schuettengruber et al. 2017). There is just one COMPASS complex in yeast, the Set1/KMT2, that is equivalent to the *Drosophila* dSet1 and the mammalian SET1A and SET1B complexes. In mammals, the SET1A/B complexes are responsible for almost the 70% of all the H3K4me3 modifications. However, there are two additional families of COMPASS-like complexes in animals including the one that comprise the *Drosophila* Trx complex and the MLL1 and MLL2 complexes in vertebrates, that also trimethylate the H3K4 residue. This second MLL complex is also found in plants and other unicellular eukaryotes, so it might have been lost secondarily in fungi. MLL1/2, usually target genes that are more related with developmental regulation and MLL2 seems to be the complex functioning on bivalent promoters. How COMPASS complexes are recruited to their target loci is still not fully understood, but it is known that in vertebrates, just like PRC1 and PRC2, they have affinity for unmethylated CpG islands. MLL1/2 contain a CxxC domain that display affinity por CpG islands, and the SET1A/B complexes are able to bind CFP1, a DNA binding protein also carrying one of those CxxC domains.

On the other hand, the remaining COMPASS complex type (Trr in *Drosophila* and MLL3/4 in vertebrates) catalyze the monomethylation of the H3K4 (H3K4me1) which in animals mainly associated to distal enhancers that are active or poised. In stark contrast with the other two complexes, MLL3/4 is only present in animals and, rather strikingly, also in the colonial choanoflagellate *Salpingoeca rosetta*, which is considered to be phylogenetically very close to the last unicellular ancestor of metazoans (Schuettengruber et al. 2017). It is not found neither in non colonial choanoflagellates nor in other holozoans such as the filasteran *Capsaspora owczarzaki*. Interestingly, MLL3/4 methylases can be recruited to target enhancers by the cell-type specific pioneer TF FoxA1, that is able

to bind to its target sites even if they are enclosed inside a nucleosome core (Jozwik et al. 2016). FoxA1 is involved in the development of the liver in vertebrates and during embryogenesis can be found bound to liver specific enhancers in endodermal cells (among them the liver precursors, Lupien et al. 2008), possibly favoring the creation of a permissive environment for the enhancer activation by other TFs. Additional TFs like GATA3 (Takaku et al. 2016) and others also display pioneer activity and it is tempting to hypothesize that they could also recruit the MLL3/4 complex in order to start sensitizing the enhancer environment, although this remains unknown (see Figure 1.3C). H3K4me1 dependent activation of enhancers happens in several ways: (i) by recruiting CBP and P300 to acetylate the H3K27, (ii) by interacting with the BAF complex, member of the SWI/SNF Tritorax complexes that include ATP dependent nucleosome remodellers that help to expose CREs out of nucleosome cores and (iii) by recruiting cohesin, a key protein complex for the 3D configuration of the genomes that mediate the interaction between distal enhancers and promoters (Local et al. 2018).

Last but not least, the activating mark H3K27ac is present both in active promoters and active enhancers, but not in poised enhancers like H3K4me1 (Creyghton et al. 2010). H3K27 acetylation is catalyzed by CBP in *Drosophila* and both by CBP and the related protein P300 in vertebrates and have been shown to be recruited to promoters and enhancers both by some pioneer TFs (Choi et al. 2016, Fuglerud et al. 2018) and by the COMPASS complexes such as the MLL1/2 and MLL3/4 cited earlier. Sadly, apart from the fact that this modification is incompatible with the H3K27me3 PRC2 repressing mark (Tie et al. 2016), the mechanisms by which the H3K27ac modification impact transcription are still poorly understood. Nevertheless, H3K27ac constitutes a specific and reliable signature to identify active enhancers genome wide (Rada-Iglesias et al. 2011).

1.2.3 The NGS revolution in the study of transcriptional regulation

Next Generation Sequencing (NGS) technologies have opened the possibility to sequence simultaneously billions of DNA fragments ranging from 50 to 200 bp in a single run and at a moderate cost. These advances have impulsed the emergence of many sequencing projects yielding an important number of genome assemblies of animals that are of great interest for the evolutionary biology. The genomes of the coelacanth (*Latimeria chalumnae*, Amemiya et al. 2013), the spotted gar (*Lepisosteus oculatus*, Braasch et al. 2016), the sea lamprey (*Petromyzon marinus*, Smith et al. 2018) or the elephant shark (*Callorhinchus milii*, Venkatesh et al. 2014) are just some examples of special interest for the evolution of vertebrates. Moreover, equally important is how it has impacted the study of transcriptional regulation. Many of the regulatory mechanisms controlling transcription presented above were known before the advent of the NGS technology. However, our understanding of transcription have greatly improved after the development of a myriad of techniques that interrogate different aspects of transcriptional regulation genome wide, and most of them are coupled to NGS as a final readout (see Figure 1.3D). RNA-seq, for example, has been used extensively to identify and quantify all the mature RNAs present in different cell populations or across different developmental stages (reviewed in Wang, Gerstein, and Snyder 2009). Recently, several single-cell versions of the RNA-seq protocols have been fully developed and these advances will allow us to assess the transcriptional state of different parts of the embryo with cellular resolution (Farrell et al. 2018, Sebé-Pedrós et al. 2018). Additionally, as we have shown before, CAGE-seq was critical

to understand how TSSs are placed within promoters and to classify those promoters in different functional categories (Carninci et al. 2006). Methylation patterns can also be assayed genome wide with MethylC-seq, that use bisulfite conversion of non methylated cytosines to uracyls coupled to sequencing to infer CpG island methylation states (Bogdanović and Lister 2017). Furthermore, although enhancers were well known before NGS development, their amount and in general the regulatory potential of the non-coding fraction of the genome (i.e. the fraction that does not harbor protein coding genes, almost the 90% of many mammalian genomes) was often overlooked or underestimated before the ENCODE project (Buttler et al. 2012) and others started to predict enhancers genome wide taking advantage of methods relying on NGS such as ChIP-seq (Johnson et al. 2007) and ATAC-seq (Buenrostro et al. 2013).

ChIP-seq (Chromatin Immuno Precipitation coupled to sequencing) experiments, for example, allow to identify all the regions in the genome that are associated with a given protein of interest (Johnson et al. 2007), using the following strategy. The protocol starts by fixating the chromatin with paraformaldehyde, which stabilizes the interactions between the different proteins and the DNA. Then, the chromatin is randomly fragmented by sonication and those fragments that are bound to the protein of interest are selected using an specific antibody. Finally, the proteins are removed and the DNA fragments are massively sequenced using NGS allowing to quantify the binding of the protein of interest to the different loci in the genome. This technique is very versatile since it allows to explore different aspects of transcriptional regulation by choosing different antibodies. It can be directed to different cell type specific TFs, to the RNAP II, to chromatin remodellers and, importantly, to epigenetic modifications of histone tails such as H3K4me1, H3K4me3, H3K27me3 or H3K27ac. The last approach, targeting histone variants, have been used extensively to predict enhancers and promoters genome wide from cnidarians (Schwaiger et al. 2014) to mammals (Visel et al. 2009, Rada-Iglesias et al. 2011). Regions that are rich in H3K4me3 and H3K27ac according to these experiments are predicted to be promoters, while regions that display H3K27ac and H3K4me1 but no H3K4me3 are probably enhancers.

The enhancer and promoter predictions using histone modifications target the nucleosomes that are flanking the exposed DNA region where TFs bind. In contrast, chromatin accessibility assays such as ATAC-seq (Assay for Transposase-Accessible Chromatin, Buenrostro et al. 2013) rely precisely on determining which regions of the DNA are exposed. The ATAC-seq protocol consist in adding to fresh chromatin a modified version of the bacterial Tn5 transposase. This transposase, in its original context, is responsible for mobilizing particular sequences of the chromosome called transposons that replicate and reinsert in different places of the genome following a cut and paste mechanism. The mechanism is hijacked in ATAC-seq for the purpose of cleaving the DNA that is exposed, given that the Tn5 is not able to cut DNA associated to nucleosomes. Then, the Tn5 also link sequencing adapters specifically to the cleaved fragments. Open chromatin regions such as promoters and enhancers (actively depleted of nucleosomes) will be much more represented in the final sequencing and can be then identified. Promisingly, single cell protocols for ATAC-seq experiments are also starting to be popularized (Buenrostro et al. 2015, Cusanovich et al. 2018).

As a summary, it is important to highlight that equivalent mechanisms of transcriptional regulation operate across most animals (e.g. the interplay between enhancers and promoters or the epigenetic modification of DNA and histone tails). In addition, the development of NGS based techniques allows to explore those mechanisms genome wide paving the way to comparative analysis among different species. Then, testing hypothesis related to the role of transcriptional changes in the evolution of animal morphology has become easier. In the next section we will explore how

the chromatin folds inside of the cell nuclei, and how other NGS based techniques such as 4C-seq and HiC have been critical to understand the role of this folding in the regulation of transcription.

1.3

Chromatin architecture and its influence in transcriptional regulation

Fully stretched, each copy of a medium sized animal genome will extend for more than 1 meter long. Strikingly, each of these copies folds to fit perfectly in a cell nucleus of only micrometers of diameter in a way that is far from random. Rather, regular patterns can be found in different cells when exploring the chromatin folding at different zoom levels depending on the technical approach employed. Then, we will use the term chromatin architecture to contrast the idea that genomes fold randomly, but without neglecting the fact that many folding patterns are dynamic as we will explore later on (specially when compared to protein structures).

First we will explore chromatin architecture at low resolution. Light microscopy based techniques such as FISH (Fluorescent In-Situ Hybridization) allow to explore the location of whole chromosomes inside the nucleus, but also the positioning of smaller loci depending of the different sets of fluorescent DNA probes used (Pinkel et al. 1988, Chambeyron and Bickmore 2004, Fabre et al. 2017, Figure 1.4A). Using FISH it was possible to determine that chromatin regions belonging to the same chromosome tend to cluster together in the nucleus conforming chromosome territories (also reviewed in Cremer and Cremer 2010). In addition, it was shown that gene rich regions are placed in the outer part of chromosome territories and are therefore more accessible to the transcription machinery. Interestingly, it was also observed that during cell differentiation some genes may switch from being hidden in the inner part of their chromosome territories to the outer part, and that these movements correlated with the onset of their transcription (Chambeyron 2005). This was one of the first clues indicating that chromosome organization might have a fundamental role in transcriptional regulation. However, only a handful of probes can be assayed simultaneously using FISH based techniques and the resolution is limited (although improving) to explore enhancer-promoter contacts efficiently. Besides that, it is known that heterochromatin regions (presented in the previous section as a compact and silent fractions of the genome) are often located towards the nuclear periphery and linked to proteins of the nuclear lamina (i.e. the inner most layer of the nuclear envelope, reviewed in Luperchio, Wong, and Reddy 2014). Then, it is possible to indirectly infer if a given genomic region is close to the nuclear periphery by exploring its association with proteins of the lamina such as different lamins and emerin, most commonly using DamID (Zullo et al. 2012, González-Aguilera et al. 2014). Those regions are called LADs (for Lamin Associated Domains) and are often transcriptionally silent.

Meanwhile, on the high resolution end, techniques such as X-ray crystallography and electron microscopy are able to get detailed deterministic pictures of small chunks of chromatin. They were extremely useful in order to determine the structure of the DNA double helix (Watson and Crick 1953) or the nucleosome cores (Richmond et al. 1997), for example. Encouragingly, in-situ ChromEMT has been used to reconstruct chromatin fiber surfaces with resolutions ranging from single nucleosomes to whole chromosomes (Ou et al. 2017, Figure 1.4B). Nevertheless, it

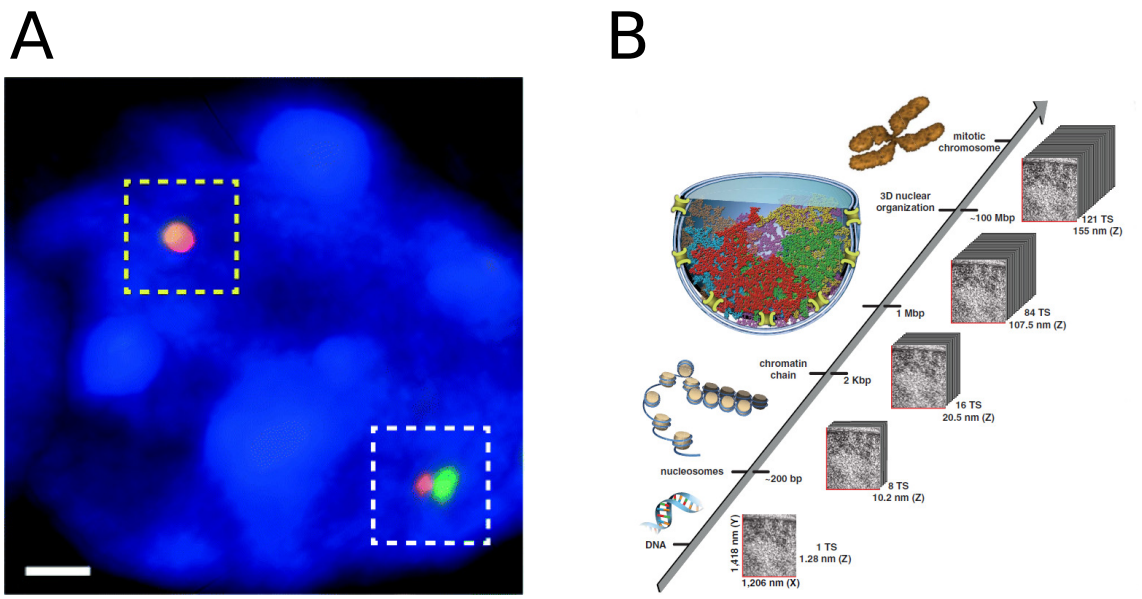


Figure 1.4: Different approaches to visualize nuclear architecture. In (A) there is a FISH experiment with two probes located in a limb enhancer (red) and its target gene *HoxD13* (green). They interact in one of the alleles but not in the other (Fabre et al. 2017, image distributed under the CC BY 4.0 license). In (B) we see how the recently developed chromEMT technique is able to reconstruct the folding of the whole genome merging electron microscopy slices, although it is not yet possible to identify specific loci (From Ou et al. 2017, reprinted with permission from AAAS).

is not yet possible to map specific loci in those reconstructions. In addition, as we commented before, nucleosome positioning can be also inferred taking advantage of ATAC-seq experiments that delineate exposed DNA fractions (Buenrostro et al. 2013). Importantly, none of these techniques are yet suited to explore efficiently the folding of DNA at the resolution that is relevant to determine enhancer-promoter contacts (from the kilobase to the megabase). In contrast, C-techniques were developed precisely to bridge the resolution gap between chromosome territories and nucleosome fibers (Marti-Renom and Mirny 2011) and have strongly influenced our way of understanding genomes.

1.3.1

C-techniques bridge the resolution gap

Chromosome conformation capture related techniques (C-techniques) constitute an instrumental approximation to infer chromatin architecture based on classical molecular biology techniques and now often coupled to NGS. The first flavor of this group of techniques was the 3C (for Chromosome Conformation Capture, Dekker 2002) and the rest of C-techniques are founded on the same principles (see Figure 1.5A). First of all, many cells are cross-linked using PFA to stabilize protein-protein and DNA-protein interactions and therefore also long range interactions like the ones involving enhancers and promoters. Then, cells are lysed to expose the chromatin and a restriction enzyme is used to cut the genome in small predictable fragments. Since DNA-protein and protein-protein interactions are still present, two restriction fragments that are relatively far away in the linear sequence might still be close if they were interacting in-vivo. Finally, ligase is added and those restriction fragments that are held close might end up ligated together forming

a chimeric molecule. The idea is that the amount of times that a chimeric molecule composed of two particular restriction fragments is found will reflect how often those restriction fragments were close in the 3D space. Both these steps and that assumption are shared by all the C-techniques. Therefore, it is important to notice that the resolution limit of C-techniques is determined by the size of the restriction fragments generated, which are often smaller than a kilobase.

Differences between different C-techniques begin when it comes to identify and quantify the chimeric molecules generated within the studied cell population (reviewed in Denker and De Laat 2016). 3C for instance relies on PCRs using primers designed in two restriction fragments representative of two loci of interest, and answers the question of whether these two particular fragments interact together by quantifying the PCR products. 3C experiments were fundamental in early studies that elucidated the interactions between the β-globin promoters and the LCR enhancers (critical for proper erytrocite differentiation, Palstra et al. 2003) and between several Irx promoters (important for the coordination of their complex regulation, Tena et al. 2011). The rest of C-techniques rely on NGS to identify and quantify chimeric molecules and as a result their scope is much bigger. 4C-seq (for Circular Chromosome Conformation Capture), for example, is designed to identify all the fragments that interact with a loci of interest (Werken et al. 2012). The full explanation of the 4C-seq protocol can be found in Materials and Methods (3.1, p.51), and the typical 4C-seq graphic representation of the results is explained in (Figure 1.5B). In contrast, HiC explore all the possible interactions between every restriction fragment in the genome, offering a general picture of the chromatin architecture of a given cell population (Lieberman-Aiden et al. 2009). Graphic representation of HiC experiments is also covered in (Figure 1.5C), and basically consists on heatmaps that represent huge square matrices (or tables) with one cell per possible pairwise contact. Each cell contains a number representing the raw or the normalized number of interactions between two given loci.

Then, HiC experiments are often the preferred alternative since it provides all the 3D information in a single assay. However, it is important to note that in C-techniques there is an important compromise between resolution and scope. Achieving high resolution genome wide requires the identification of many chimeric molecules, and that means a lot of sequencing depth. Then, if high resolution around a specific locus is needed it might be more reasonable to choose an alternative. 4C-seq is one of them, but there are also other possibilities such as 5C (Dostie et al. 2006) and Capture-C (Jäger et al. 2015) that offer a wider look over a locus of interest. Capture-C, for instance, is based on selecting chimeric molecules involving restriction fragments of interest using probes before sequencing them in a way that is equivalent to HiC. Usual graphic representations of Capture-C and 5C experiments are mostly equivalent to those of HiC. Of especial interest is the development of hybrid C-techniques that include the selection of a subset of the interactions based on antibodies such as ChIA-PET (Fullwood et al. 2009) and most recently HiChIP (Mumbach et al. 2016). HiChIP has allowed, for example, to study which contacts are driven by specific architectural proteins such as cohesin or CTCF (discussed later) or are related with different chromatin states using histone tail modification antibodies such as H3K27ac or H3K27me3 (Rowley et al. 2017). A detailed protocol of HiChIP using histone mark antibodies can be also found in Material and Methods (see 3.2.1, p.62).

Lastly, it is necessary to explain one of the main limitations inherent to the protocols of C-techniques (see Figure 1.5D). That is the fact that C-techniques are only able to identify pairwise contacts since, per cell, only two ligation events per restriction fragment can be captured (one per homologous chromosome). That holds true regardless of the number of simultaneous interactions

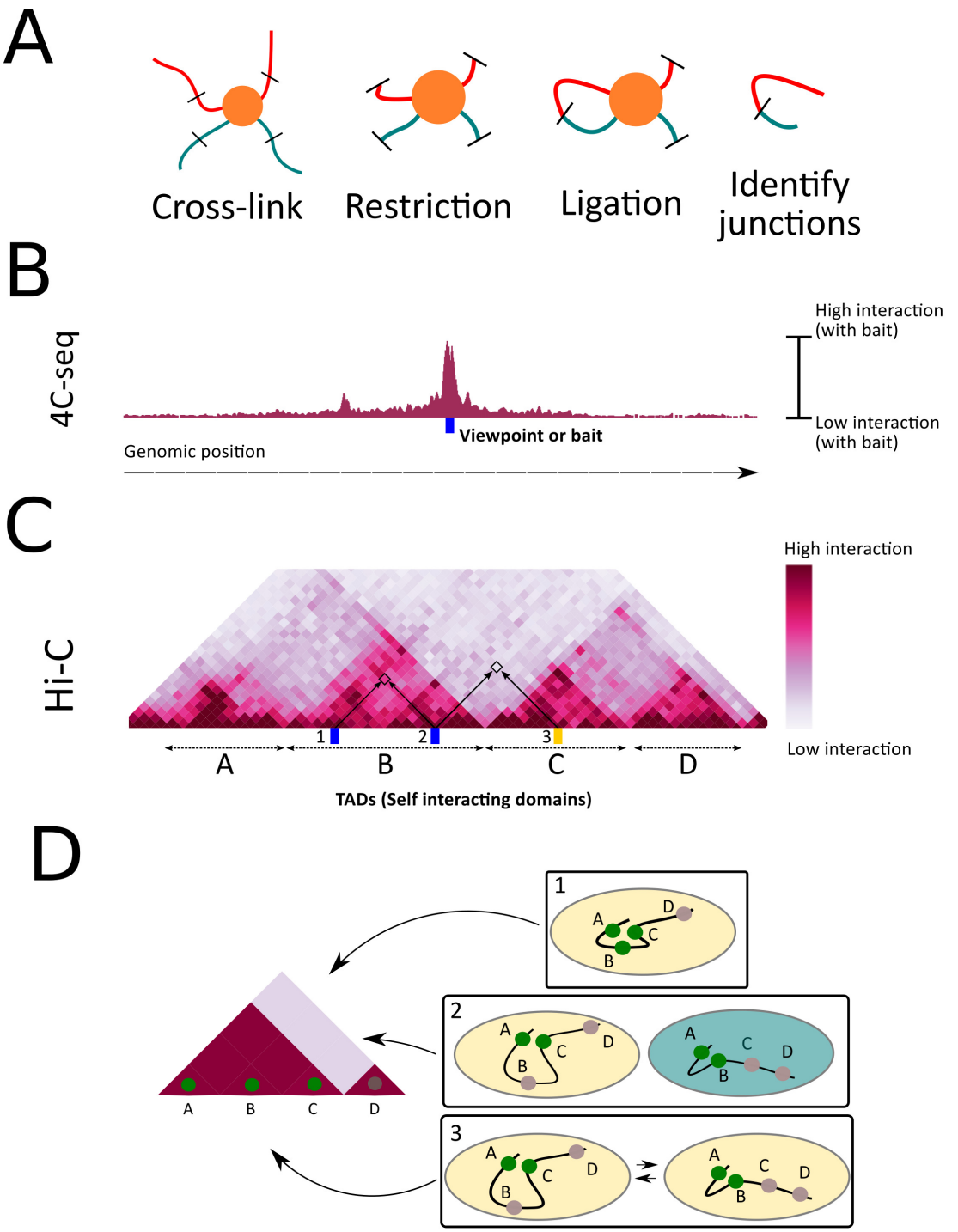


Figure 1.5: Principles and limitations of C-techniques. (A) Four basic steps that are common to all C-techniques. DNA-protein and protein-protein interactions are fixed with PFA, then the cross-linked chromatin is digested with a restriction enzyme and religated so that distant restriction fragments in the linear sequence but close in the 3D space might become consecutive. Finally, ligation junctions are identified and quantified. (B) Typical 4C-seq plot, technique that explore the interactions between a given bait (blue rectangle) and the rest of genomic loci. The genomic position is represented in the x-axis while the frequency of interactions with the bait is represented in the y-axis. (C) Typical HiC heatmap representing the matrix of interactions. The frequency of interactions between each pair of loci can be assessed by looking at the square located at their intersection. Darker colors represented higher frequency of interactions and then loci 1 and 2 interact often while loci 2 and 3 interact seldom. The four dark triangles hint the presence of TADs. This will be explained further later. (D) One of the main limitations of the C-techniques is that it only samples one interaction per allele. Therefore, the heatmap on the left is not able to discriminate between the three scenarios on the right: (1) Loci A, B and C always interact together; (2) in different cells, the A locus interact either with locus B or with locus C or (3) in different moments but in the same cell locus A switches and sometimes contact locus B and sometimes contact locus C. Indeed, infinite intermediate scenarios are possible.

involving a specific locus and even if the technique is performed with single-cell resolution (Nagano et al. 2017). Therefore, to get a more informative picture, C-experiments are often performed in big cell populations. Then, imagine some HiC data informing that locus A, B and C interact strongly between themselves. This observation can be equally well explained by several different scenarios: (i) the three locus interact together in every cell, (ii) A either interact with B or with C, or B interacts with C, but the ternary complex is never found and (iii) any intermediate scenario that compensate the final interaction readout. Excitingly, super-resolution microscopy might become soon ready to solve some of these debates (Bintu et al. 2018). In any case, if A is a promoter it is clear that it could interact with enhancers placed both in B and in C and then this moderate level of uncertainty might not be critical. In fact, as we will explore now, C-techniques have been unquestionably useful to discover and describe several patterns of chromatin folding crucial for transcriptional regulation: compartments, TADs and loops.

1.3.2

A/B compartments

We will proceed from big to small folding patterns, and then we will focus first on the partition of chromosomes in A and B compartments. A and B compartments were described using HiC almost ten years ago when the maximum resolution that could be achieved was 1 megabase (Lieberman-Aiden et al. 2009)). Note that this resolution was still not informative for enhancer-promoter interactions. However, an interesting pattern arises when looking at the heatmaps of contacts of whole chromosomes at 1Mb resolution. The first obvious observation is that the contacts around the main diagonal are really strong (represented by dark colors), which is expected since loci that are close in the linear sequence of DNA tend to also be close in the 3D space. However, looking a bit further away, a chessboard pattern arises with matching darker and lighter squares, suggesting the presence of two big compartments in the chromosomes (Figure 1.6B and 1.6D). The chessboard pattern is even more apparent if we transform the original matrix using an observed/expected correction by distance or a row/column Pearson correlation (Figure 1.6C and 1.6E). Principal Component Analysis also readily identify two compartments in HiC matrices.

Interestingly, each compartment is strongly enriched in different epigenetic marks. Compartment A is the one enriched in active epigenetic marks like H3K36me3 and display more regions of accessible DNA, more coding genes and higher levels of expression of these genes. Meanwhile, Compartment B is gene poor and depleted of active epigenetic marks (Figure 1.6A). In addition, it has also been shown that Compartment B regions are also more densely packed using FISH. Furthermore, switching of big genomic regions from Compartment B to Compartment A during cell differentiation coupled to gene activation have been described (Dixon et al. 2015). Although not fully explored, this is highly reminiscent of active genes been actively placed towards the outer part of chromosome territories (Chambeyron 2005) and compatible with the old hypothesis of transcription factories that propose that RNAP II cluster together with active genes in particular nuclear locations (Iborra et al. 1996). Importantly, the development of the HiC technique allowing to explore genomic interaction with 1kb resolution has refined very much our understanding of compartments (Rao et al. 2014). It has been proven that the big genomic chunks of several megabases originally assigned either to the A or to the B compartment can be further subdivided in smaller regions that escape the general gross trend and switch from A to B and vice versa. In

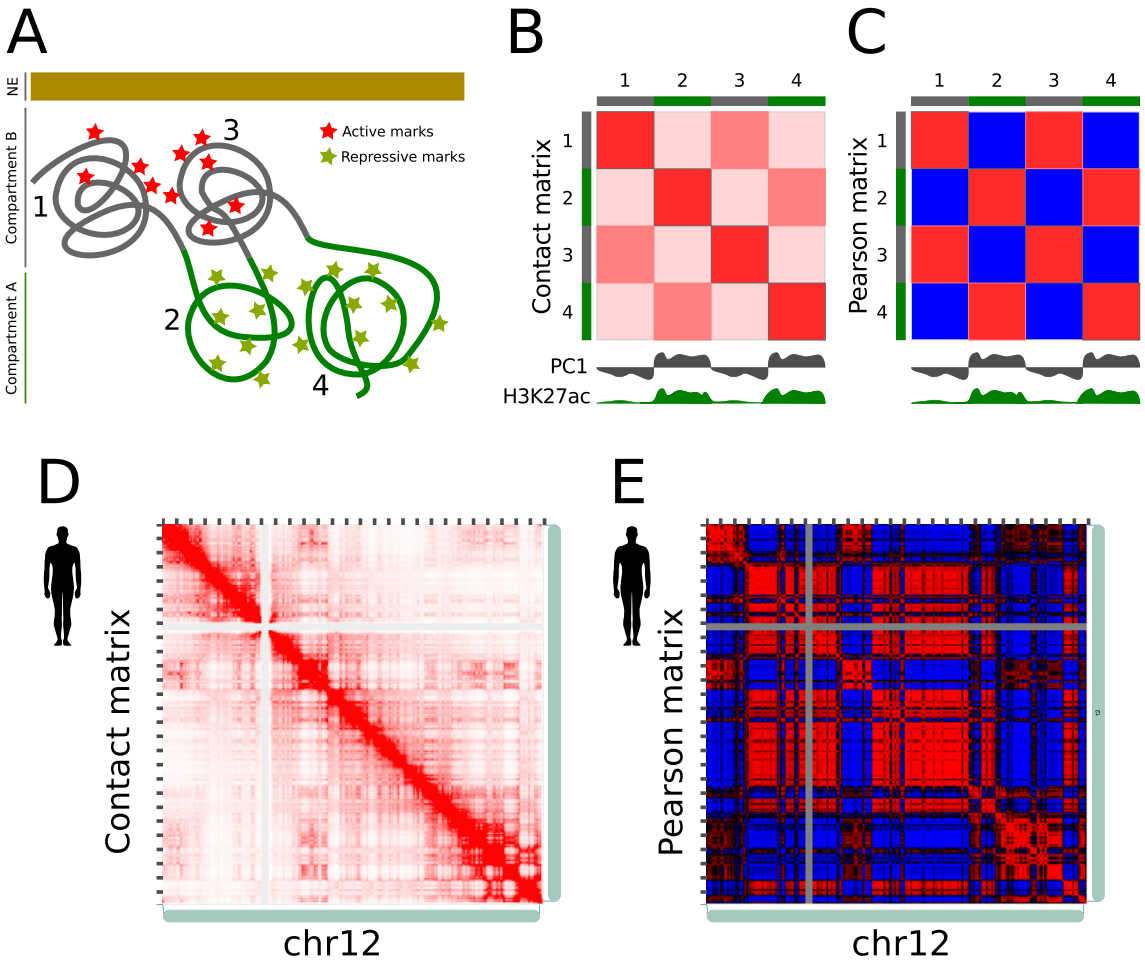


Figure 1.6: Compartment identification in HiC matrices. (A) Epigenetically active and inactive regions cluster together in space. (B) Cartoon of a simplified HiC matrix showing the two compartments. The chessboard pattern allows to classify the loci 1 and 3 in the compartment B and the loci 2 and 4 in the compartment A. They can be identified by the first principal component (PC1) and compartment A is established according to the presence of active epigenetic marks. (C) The pearson transformation of the data allows to enhance the compartment classification. In (D) and in (E) we reproduce the actual raw and pearson transformed HiC matrices of human GM12878 cells. The heatmaps were browsed using *Juicebox* (Durand et al. 2016) and the experiments are from Rao et al. 2014.

addition, it has been shown that the A compartment can be also further subdivided in A1 and A2 and the B compartments in B1 to B4. Perhaps the most relevant division is the one separating the B1 compartment from the rest, because this B1 compartment is heavily enriched in the H3K27me3 epigenetic modification, associated with facultative repression of transcription mediated by polycomb complexes. In contrast, B2 to B4 compartments are more related to constitutive heterochromatin marks.

1.3.3 Topologically associated domains (TADs)

In order to identify the next feature, Topologically Associated Domains (TADs), it is also necessary to use C-techniques that are able to sample multiple combinations of pairwise contacts (e.g. HiC

or 5C) but with higher resolution than the resolution needed to detect compartments. They were first described in 2012 in *Drosophila* using HiC reaching 10kb resolution although they were termed Physical Domains (Sexton et al. 2012). When *Drosophila* HiC matrices were examined, it became apparent that the genome was partitioned in many domains of up to several kilobases in length. Interactions between loci belonging to the same domain were highly frequent, however, interactions between loci belonging to different domains seemed to be disfavored. This is readily visible in the HiC heatmaps as a succession of dark squares overlapping the diagonal. Later in 2012, HiC and 5C experiments performed in some human and murine cell lines and tissues confirmed the global segmentation of mammalian genomes in what was called TADs (Nora et al. 2012, Dixon et al. 2012). In this case, interestingly, mammalian TADs could reach sizes of more than 1Mb which fitted extraordinarily well with the examples of the largest RLs that had been described so far. For example, both *Shh* and its ZRS enhancer belonged to the same TAD despite the distance of 1Mb separating them (Symmons et al. 2016, Figure 1.7A). In addition, the two RLs that were proposed to regulate the HoxD cluster in mouse were also embedded in two abutting TADs with the genes located precisely at the boundary between the two TADs (Andrey et al. 2013). Therefore, it was proposed that enhancers and promoters needed to belong to the same TAD in order to interact, and that elements present at TAD boundaries prevented spurious interactions between enhancers and unintended core promoters across different TADs (Figure 1.7B).

Importantly, the role of TADs both in favoring and preventing enhancer-promoter interactions has been extensively tested functionally. An example of the latter are the experiments by Lupiañez et al. 2015 studying both the murine and the human *Ihh/Epha4/Pax3* loci. Chromosomal rearrangements such as inversions, deletions and duplications around the human locus were associated with a variety of severe limb malformations. First of all, they showed that the WT 3D configuration of the locus was conserved between humans and mice. This 3D configuration consists in three TADs, each of the TADs containing one of the three genes plus a set of enhancers. Out of the three genes only *Epha4* was expressed in limbs, and suspiciously, all the rearrangements related to limb malformations encompassed TAD boundaries separating either the *Ihh* TAD or the *Pax3* TAD from the *Epha4* TAD (that presumably contains limb specific enhancers). Engineering equivalent mutations to those observed in humans in mice they showed that indeed either *Ihh* or *Pax3* were able to interact with genomic regions belonging to the *Epha4* TAD. Furthermore, in contrast to what happens in the WT condition, expression of *Pax3* or *Ihh* could be detected in embryonic limb buds in equivalent territories to those belonging to the expression pattern of *Epha4*. This functionally proves that TAD boundaries surrounding the *Epha4* RL are needed in order to prevent other genes to hijack *Epha4* enhancers, get expressed in limbs and distort their normal development. But apart from preventing undesired interactions, TADs also seem to facilitate contacts between distant promoters and enhancers. That is the case of the ZRS enhancer for example, that drives the expression of *Shh* to the posterior compartment of the limb bud despite being more than 1Mb away. Big inversions were also engineered in mice by Symmons et al. 2016 placing the ZRS enhancer at half the distance to the *Shh* promoter. However, in this new configuration, the ZRS was located beyond the TAD boundary and that genotype resulted in the absence of *Shh* in the limb bud accompanied by severe limb phenotypes resembling those caused by the full deletion of the enhancer.

The development of high resolution HiC protocols coupled to several loss of function experiments targeting different architectural proteins are starting to unravel the mechanisms by which TADs are formed and maintained. In mammals, the so called extrusion model is now widely accepted as the

most probable mechanism forming TADs (Sanborn et al. 2015)). It mainly involves the interplay between two extruding rings, role that is likely played by a protein complex called cohesin, and several brakes located at TAD boundaries that are originated by CTCF dimers. This is supported by the different TAD alterations observed upon the depletion of CTCF (Nora et al. 2017), cohesin (Rao et al. 2017), cohesin loaders (Schwarzer et al. 2017) or unloaders (Haarhuis et al. 2017). First, cohesin rings load up at a certain region of the chromatin in such a way that the DNA fiber end up threaded inside both rings. Then, cohesin rings start to slide through the DNA fiber in opposite directions, bringing together pieces of chromatin located progressively further and further away. Meanwhile, CTCF is a zinc finger DNA binding protein that recognize a long asymmetric motif that is commonly found at TAD boundaries in divergent orientations. That is so because two CTCF proteins bound to convergent CTCF binding sites are able to form dimers and bring together distant loci (Rao et al. 2014, Gómez-Marín et al. 2015), and these loops are able to stop cohesin rings from extruding. Cohesin rings stopped at CTCF dimers are then unloaded from the chromatin by the protein WAPL (Haarhuis et al. 2017). In *Drosophila*, however, chromatin states seem to play a much more prominent role in delineating TADs than architectural proteins, although the role of transcription in this process is still controversial (Rowley et al. 2017). In any case, GRO-seq (assessing nascent transcription) and RNAP II occupancy are by far the best predictors of boundary locations in *Drosophila*. It has been proposed that these small active regions form small active A compartments in between two B compartments, and that A-A compartment interactions with other boundaries drive the insulation of domains. Intriguingly, *Drosophila* display a large amount of architectural proteins including CTCF, but they seem to be only required for the insulation of specific loci (Hou et al. 2012). Strikingly, CTCF in *Drosophila* does not seem to form dimers when bound to convergent binding sites as it has been described for mammals (Rowley et al. 2017). Interestingly, actively transcribed regions are also enriched in mammalian boundaries and some of them are sufficient to establish insulation between two domains, speaking of a conserved role of active chromatin in the formation of TAD boundaries.

An indirect measurement of TADs importance in gene regulation is their degree of conservation. First of all, TADs as a genomic feature seem to be conserved among many bilaterians of divergent groups. HiC experiments have proven their existence extensively in more than 50 mammalian species (Vietri Rudan et al. 2015, the DNA zoo project: Dudchenko et al. 2017) plus in chicken (Gibcus et al. 2018), zebrafish (Kaaij et al. 2018), mosquitoes (Dudchenko et al. 2017) and flies (Sexton et al. 2012). Shockingly, they seem to have been lost secondarily in the nematode *Caenorhabidtis elegans*, where such a compartmentalization is just observed in the female X chromosomes and does not seem to be involved in constraining enhancer-promoter interactions (Crane et al. 2015). Despite the sparse collection of organisms with HiC experiments available there are other lines of evidence pointing towards a widespread presence of TADs, at least among bilaterians. One of them is that CTCF is a bilaterian novelty present in the majority of bilaterian phyla whose main function is the insulation of TADs (Heger et al. 2012). Interestingly, *C. elegans* and some other closely related nematodes lost CTCF secondarily (Heger, Marin, and Schierenberg 2009) together with the TAD organization. Another indirect proof is the extensive conservation of microsynteny across distantly related phyla (Irimia et al. 2012). Microsynteny is the precise ordering of a group of genes in a given locus, and it has been shown to be often related to the presence of distal enhancers. If two genes are kept together in highly distant lineages it is likely that the genomic region may not be reshuffled because distal enhancers of one of the genes are located in introns of the other gene or even in the genomic region beyond that other gene. Then,

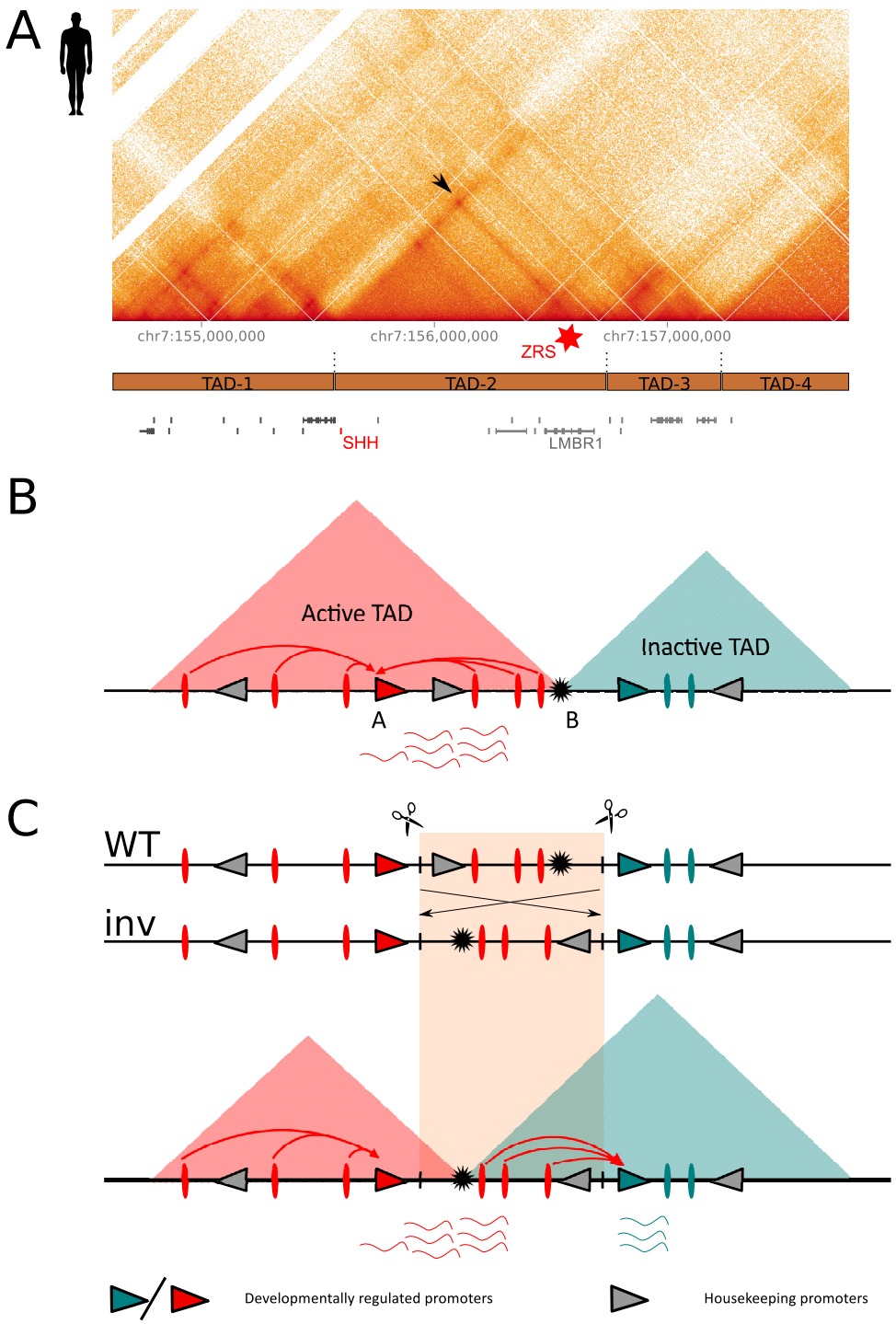


Figure 1.7: Topologically Associated Domains (TADs) are critical to connect enhancers with target promoters. (A) TAD structure around the human *Shh* locus in GM12878 cells. The arrow pinpoints the loop (seen as a dark dot) connecting *Shh* with its limb enhancer ZRS. Heatmaps from Rao et al. 2014 visualized using *Higlass* (Kerpedjiev et al. 2018) (B) Current model by which enhancers are only able to activate promoters located within the same TAD. That is the reason why in the cartoon the gene A is active but B is kept silent. If an inversion happen with breakpoints that encompass the TAD boundary (C) some former enhancers of the gene A TAD could switch to the gene B TAD activating gene B ectopically.

separating those two genes by a genomic rearrangement would mean the loss of interactions between one of the genes and their distal enhancers, and that is counter selected. Accordingly, *C. elegans* display a much less microsyntenic regions than other bilaterians, further indicating that microsynteny conservation might be a good proxy for the presence of distal enhancers and TADs.

Apart from the fact that the segmentation of genomes in TADs is conserved, precise locations of TAD boundaries across different cell types and even between different species is also the general trend. For example, 75.9% of murine TAD boundaries are also TAD boundaries in equivalent human loci (Dixon et al. 2012)). Furthermore, no genomic rearrangement happening between distant mammalian species such as mouse and dog distorted TAD structures, since all the breakpoints were precisely allocated at TAD boundaries (Vietri Rudan et al. 2015). Accordingly, the limits of arrays of conserved non-coding elements (CNE, many of them enhancers) between chicken and mouse predict the positioning of TAD boundaries, further indicating that the location of the boundaries is evolutionary constrained (Harmston et al. 2017). Finally, probably the most extreme example of deep conservation of a TAD boundary described so far involve the Six gene clusters. Six genes are transcription factors displaying complex expression patterns during development, generally found either in pairs or in groups of three since the eumetazoan ancestor. Strikingly, a TAD boundary bisecting the Six gene loci in two is conserved at least from echinoderms to vertebrates (Gómez-Marín et al. 2015). The separation of the Six loci in two RLs results in almost non overlapping expression domains of the Six genes situated at each side of the boundary.

Interestingly, it has been proposed that the precise allocation of genomic rearrangement breakpoints around TAD boundaries is not just caused by the fact that the alteration of TADs is heavily counter selected. Rather, it has been shown that the extrusion of the chromatin generate topological stress, and that the topoisomerase TOP2B that produce double strand breaks to relief that stress is enriched at TAD boundaries, which seems to be mediated by interactions with cohesin and CTCF (Canela et al. 2017). Therefore, TAD boundaries might be hotspots for chromosomal reorganizations per se due to the higher frequency of double strand breaks events happening around. However, it is important to bear in mind that several disease associated examples of TAD distortion caused by chromosome rearrangements occurring far from TAD boundaries have been described. For instance, the already presented example of the *Ihh/Epha4/Pax3* locus (Lupiañez et al. 2015, Figure 1.7C). Then, it is tempting to speculate that TAD reorganization might be an evolutionary mechanism producing drastic transcriptional changes and perhaps leading to the sudden appearance of morphological innovations.

1.3.4 Intra-TAD loops

We will discuss now the looping events happening within TADs. Some of them, like the looping between the β-globin locus and the LCR enhancers, were known long before TADs were discovered. However, now we know that their interactions are favored because they belong to the same TAD and our understanding on how do they form is benefiting from the development of high resolution HiC. In HiC matrices, loops can be spotted as dark dots that connect to loci that are surrounded by a light region (see Figure 1.7A) and have been extensively identified in several human cell lines (Rao et al. 2014).

Several proteins have been found associated to chromatin loops, and not surprisingly, the most

enriched were CTCF and cohesin. Reassuringly, CTCF enrichment in chromatin loops is also associated with a convergent pattern of its motif at each of the sides. Importantly, CTCF mediated looping occurs at TAD boundaries but also internally, producing slight variations of the intra-TAD folding. However, intra-TAD CTCF binding is often weaker and, together with internal TAD structures, much more variable even within different mammals (Vietri Rudan et al. 2015). Some of these CTCF bindings are important to mediate enhancer-promoter contacts, although CTCF looping does not seem to be the main mechanism bringing together these elements. In contrast, other proteins such as YY1 seem to play a more relevant role in bringing together promoters and enhancers (Weintraub et al. 2017). This protein have been also shown to form dimers as the main mechanism to bring enhancers and distant promoters together, which is reminiscent of the behavior of CTCF. Furthermore, like CTCF, this protein is also ubiquitously expressed. In addition, ZNF143 seems to be also involved in the establishment of enhancer promoter loops from the promoter side (Bailey et al. 2015). Lastly, it has been already cited that the H3K4me1 histone modification seem to be crucial for the stabilization of cohesin complexes around enhancers (Local et al. 2018). However, the relationships between those many factors and cell type specific TFs are still poorly understood.

Finally, it is important to comment that sometimes the appearance of cell type specific enhancer-promoter loops is linked to the onset of transcription. This was reported for the β-globin locus in murine erythrocites (Deng et al. 2012) and it is now also supported by genome wide analysis in human cell lines Rao et al. 2014. However, this does not seem to be the general trend. In contrast, constitutive or at least anticipated interactions between promoters and their target enhancers are more widely observed (Tena et al. 2011, Montavon et al. 2011, Symmons et al. 2016).

1.4

Developmental gene regulatory networks and the evolution of body plans

So far, we have discussed that transcriptional regulation is fundamental during development in order to organize the positioning of different cell types within the embryo and also in instructing some key morphogenetic events. In addition, we have presented common principles underlying animal transcriptional regulation such as the interplay between promoters, enhancers and cell type specific TFs and how epigenetic changes and the 3D architecture of the nucleus add additional yet crucial layers of control. Now that we have introduced these common rules, we need to focus on how combination of signaling molecules and specific TFs interact in order to robustly deploy different transcriptional programs in different parts of the embryo. Unraveling these interactions is an arduous task since computational approaches predict that there are more than two thousand different TFs both in the human and in the mouse genomes (Fulton et al. 2009). The concept of transcriptional gene regulatory network (GRN) provide an instrumental theoretical framework both to structure our current knowledge of regulatory interactions and to predict the outcome of different perturbations to the developmental system (e.g. mutations affecting different proteins or CREs). The model of the GRN represents the regulatory interactions happening during development as a huge network with genes located at the nodes and connected by either enabling or inhibitory

interactions. Even though our knowledge of the nodes and interactions of the networks is still limited, the thorough study of many circuits and subcircuits have made it possible to establish some useful generalizations about GRN topologies and hierarchies.

1.4.1 Two different views on the structure of GRNs: kernels and ChINs

Erwin and Davidson proposed and described 3 central elements composing developmental GRNs: (i) kernels, (ii) I/O switches and (iii) terminal gene batteries (reviewed in Davidson and Erwin 2006). First of all, kernels are sub-circuits placed at the top of the hierarchy of GRNs and their role is to specify the fate of a particular cell population that will give rise to a particular structure of the body. They do so through the activation of a much larger amount of downstream circuits, and in turn, effector genes responsible for the phenotype and behavior of the cells. Kernel circuits are composed of highly conserved sets of TFs that are heavily interconnected between themselves and that are often involved in self regulatory feed forward loops that ensure, upon activation, the robust deployment of the kernel transcriptional program. Importantly, kernels tend to be highly conserved during evolution due to their position in the GRN hierarchy. Indeed, the extreme conservation of the heart kernel from mammals to *Drosophila* and perhaps even cnidarians (Wijesena, Simmons, and Martindale 2017) is one of the most striking examples (Figure 1.8A). Secondly, I/O switches might be also composed of conserved circuits but they are often redeployed more easily for different purposes in disparate parts of the embryo. Many I/O switches provide positional information, and they do so because they are often composed by well known cell signaling pathways such as Hedgehog, Wnt, TGF-beta (including BMPs), FGFs, etc. These pathways comprise morphogens that are able to diffuse along the embryo in a way that their concentration convey spatial information. They are also composed of the receptors of those morphogens and of several transducing factors that interpret the morphogen signaling and in turn activate target cellular processes or specific TF circuits (Dominguez-Cejudo and Casares 2015). Importantly, I/O switches may have functions at different levels of the GRN hierarchy, from activating kernels to repressing peripheral sub-circuits that are closer to terminal gene batteries. Finally, gene batteries are the terminal targets of the GRN and encode proteins involved in the cellular processes that define the different characteristics and the behavior of the cells. For example, the contractile proteins or the ion channels present in the cells of the heart will belong to this category (Martinson et al. 2014).

An alternative set of categories for GRNs was proposed by Gunter Wagner, including (i) positional information signals, (ii) character identity networks (ChINs) and (iii) realizer genes (Wagner 2007). They largely overlap with switches, kernels and batteries respectively. However, the ChIN and kernel concepts are somehow divergent despite the fact that both refer to regulatory circuits that self-sustain and instruct specific populations of cells to become a specific organ or character, instead of any other. In the definition of kernel is implicit that they are ancient circuits devoted to the patterning of highly conserved structures. In contrast, the definition of ChIN relies heavily in the concept of character identity that is based on homology: if two characters from different species evolved from a common ancestral character they are considered to be homologous and therefore have the same character identity. Then, they also share an homologous ChIN as well. Importantly, every individual character must have an individual ChIN associated, regardless of how old the character is. It is also worth noting that, despite homologous characters share ChINs, differences in

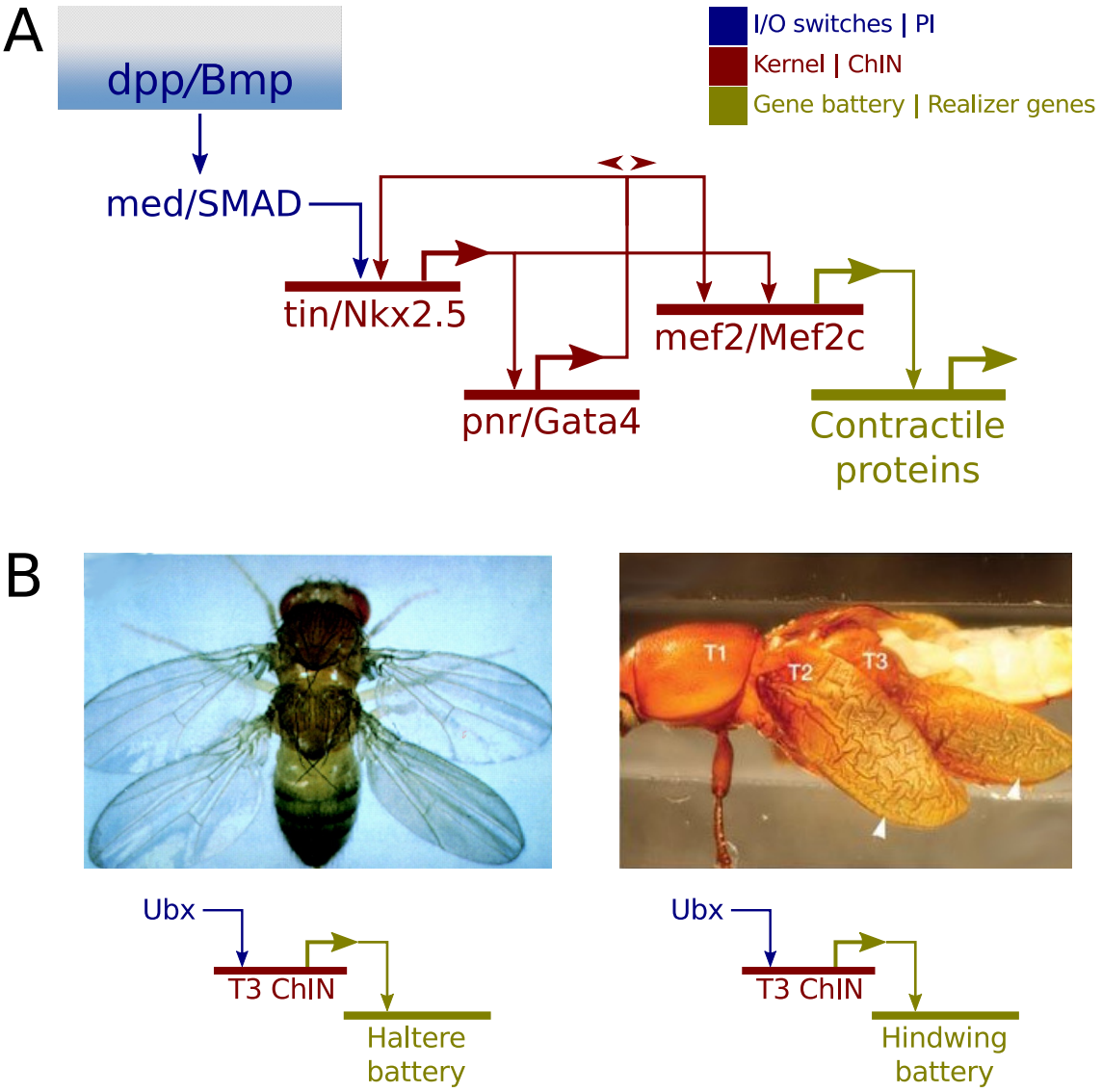


Figure 1.8: Examples of Gene Regulatory Networks. (A) Some of the conserved circuitry between vertebrates and *Drosophila* of the heart field network. (B) Homologous kernels or ChINs can instruct the formation of homologous yet quite divergent structures. Knocking down *Ubx* in *Drosophila* leads to a T3/T2 segment transformation turning halteres into wings (left image, from Weatherbee et al. 1998, distributed under the CC BY 4.0 license). However, in the case of the beetles, the same transformation turns the hindwings in a second pair of elytra (right image, from Tomoyasu, Wheeler, and Denell 2005, reprinted with permission from Springer Nature).

the genes located downstream might generate important variations in the final phenotype between different species. This is beautifully exemplified by the opposite effects of the transformation of the identity of the thoracic segment 3 (T3) in thoracic segment 2 (T2) between flies and beetles (Figure 1.8B, Tomoyasu, Wheeler, and Denell 2005). Regardless of our preferred terms, if we assume the GRN framework, then inheritable changes in development are the results of the evolution GRN circuitry. We will now explore how this framework can be useful in order to predict and test how different types of perturbations can alter the final developmental output.

1.4.2

GRN topologies influence the evolution of transcriptional regulation

First of all we will consider the consequences of tinkering GRNs at different hierarchical levels. The most accepted view is that changes happening at the periphery of the GRN, close to the terminal gene batteries, generate more local and subtle changes that are in general more likely to be tolerated. This kind of changes occur frequently and are observable even between closely related species, and one of the classical examples are the many variation in the wing pigmentation among different *Drosophila* species due to differences in the regulation of *yellow* (Figure 1.9A, Gompel et al. 2005). On the other hand, modifying nodes and connections from central parts of the GRN such as kernels seem to be heavily counter selected. On one hand because they are largely conserved and on the other because distorting some of their elements experimentally often leads to lethality due to the loss of whole parts of the embryo. That is the case of the loss of Notch signalling that results in the absence of most mesodermal derivatives in sea urchin embryos (Sherwood and McClay 1999). An intermediate case are the I/O switches or positional information signals. In general, their internal circuits are conserved but their modules can be reused to pattern different cell populations within the embryo. Although less common, changes at these levels of the GRN potentially lead to more drastic morphological changes than tickling with terminal batteries. A recent illustrative example of this is the probable co-option of *Shh* to the patterning of vertebrate paired-appendages (Figure 1.9B, Letelier et al. 2018a). This gene transitioned from participating in the patterning of dorsal fins in the vertebrate ancestor and in extant agnathes (e.g. lampreys and hagfishes) to the patterning of pectoral fins, a novelty of gnathostomes (the group comprising the rest of vertebrates). Interestingly, distinct yet slightly similar structures such as vertebrate appendages, insect wings and cephalopod tentacles seem to be patterned by strikingly similar circuits (Tarazona et al. 2018). However, those characters are far from homologous since there was not a common ancestral character they evolved from. Instead, this phenomenon known as deep homology seems to hint that modularity of GRNs sometimes facilitate the rapid appearance of novelties by reusing the previously assembled circuitry of regulatory genes and their downstream targets (recently reviewed in Tschopp and Tabin 2017). Lastly, it is also important to note that quantitative differences in gene expression instead of qualitative changes in GRN connectivity have been also related to morphological variation. For instance, the relationship between the levels of *Bmp4* and *CaM* and the different shapes of Darwin finches beaks (Mallarino et al. 2011).

Secondly, we will discuss which mutational mechanisms are potentially able to alter GRNs and generate transcriptional novelty. In the context of GRNs the binding of TFs to enhancers to activate or repress a given gene, together with other mechanisms such as protein protein interactions, are responsible to make connections between the different nodes. Therefore, it has been largely proposed that mutations in the non-coding fraction of the genome, mainly affecting cis-regulatory elements like enhancers, are the main drivers of evolutionary novelty (reviewed in Carroll 2008). There are several arguments that support the importance of non-coding mutations in the evolution of gene regulation. One of them relies on the fact that the same circuits and the same genes are reused for several functions in different parts of the embryo and in different developmental stages. Then, eliminating or modifying a coding gene involved in many different processes will most likely be deleterious since it will affect many developmental processes instead of one. In

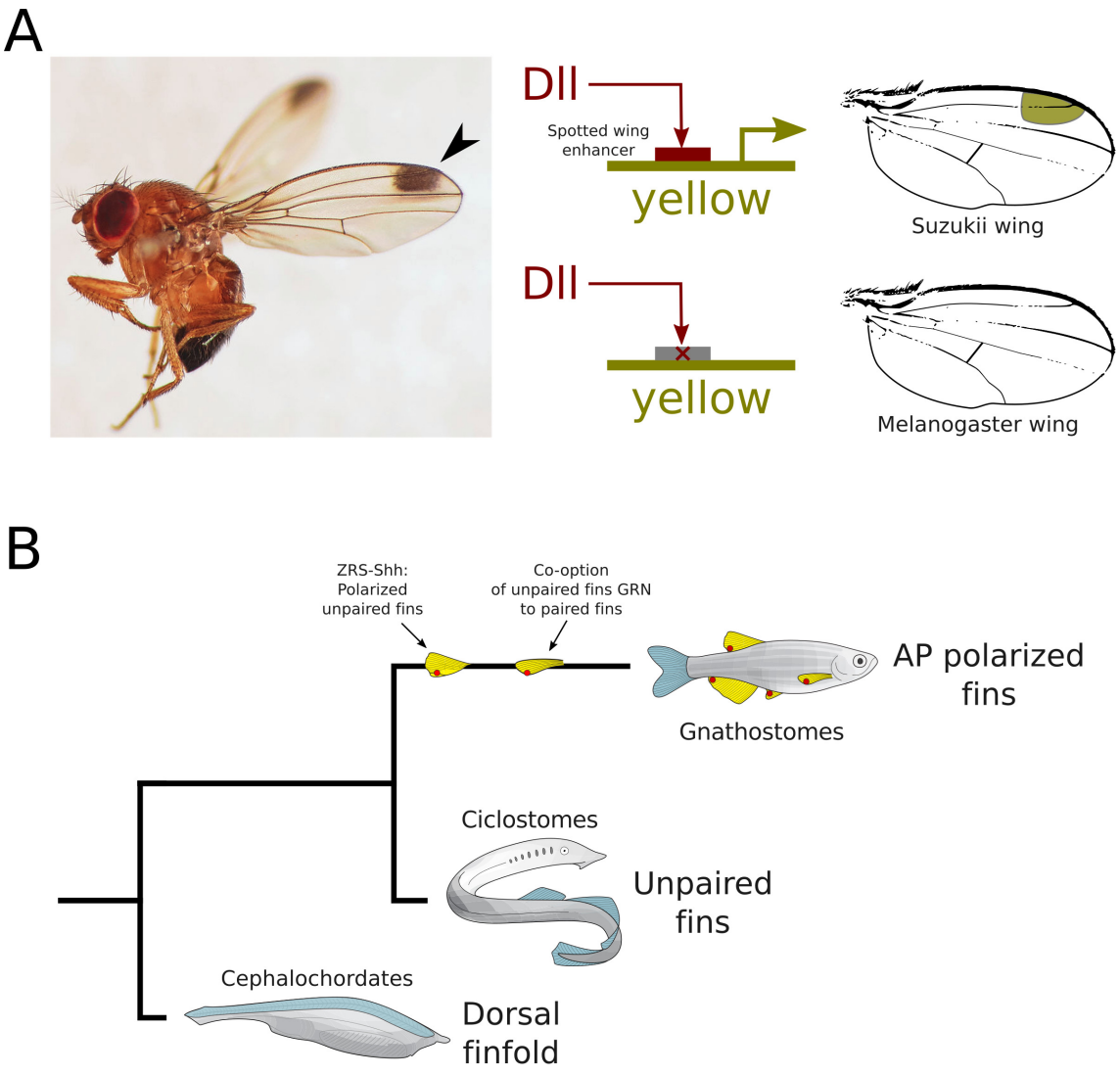


Figure 1.9: Two different ways of evolving GRNs. (A) Classic example of the evolution of GRNs through the modification of enhancers of peripheral genes of the network. The wing pigmentation of *Drosophila suzukii* depends upon the expression of the gene *yellow* driven by a specific *Dll* enhancer. This enhancer is mutated in *Drosophila melanogaster* and the wing pigmentation is absent (Prud'homme et al. 2006). Left image from Martin Cooper, distributed under the CC BY 2.0 license. (B) The co-option of most of the GRN responsible of the patterning of unpaired fins to the lateral plate mesoderm is the most plausible mechanism for the appearance of paired appendages in the gnathostome lineage (Letelier et al. 2018a).

contrast, enhancers are much more modular entities that in many cases only direct the expression of a particular gene to a particular place. In addition, the finding of a very similar toolkit of developmental proteins in all animals, including cnidarians, further argued that the coding fraction of the genomes had remained mostly unchanged (reviewed in Rokas 2008). The experiments that showed that the human *Pax6* gene was able to generate ectopic compound eyes in *Drosophila* like its *Drosophila* counterpart *eyeless* (Halder, Callaerts, and Gehring 1995) and vice versa, with *eyeless* inducing ectopic expression of eye related genes in *Xenopus* embryos (Onuma et al. 2002), also helped to reinforce this view. In addition, many examples of evolution of morphological characters linked to the evolution of enhancers appeared in the literature. However, although the importance of the evolution of the non-coding genome is difficult to deny, some fair criticism have questioned

the excessive focus on enhancers as the almost unique way of evolving gene regulation.

Alternative and plausible mechanisms for evolving GRNs include the asymmetric divergence of the sequence of paralog genes originated from duplication events (both in the case of local tandem duplication or whole genome duplication events), or the acquisition of new domains by TFs (reviewed in Holland et al. 2017 and Lynch and Wagner 2008 respectively). The appearance of a novel PRD-class homeobox cluster of TFs in mammals through tandem duplication is one example of the former (Maeso et al. 2016). The acquisition of a new aminoacidic domain by *Ubx* in insects, that turns out in the suppression of the legs of the abdominal segments is one of the latter (Galant and Carroll 2002). Here, we will mainly focus on the evolution of the non-coding genome, but without intending to undermine complementary evolutionary mechanisms like the ones mentioned above. In particular, we will try to address if changes in the 3D architecture of the genome may drive relevant regulatory novelty by rewiring enhancer and promoter connectivity in developmental GRNs.

1.4.3

Extreme GRN conservation underlying body plans stability

We have previously introduced that the fossil record indicates that the body plans of all extant organisms appeared rapidly during the Cambrian explosion (560 mya). This poses one of the most exciting enigmas in evolutionary biology, since it entails the sudden and early appearance of a great proportion of the novelties that are in turn responsible for the morphological differences found between divergent animal groups such as insects and vertebrates (Erwin et al. 2012). Consequently, it also requires a posterior slowing-down in the pace of the appearance of novelties once those body plans appeared. But to explore this further we need to clarify first what do we understand as a body plan, that is a concept that relied mostly in comparative anatomy and embryology (as discussed in Slack, Holland, and Graham 1993; Willmore 2012). A given body plan can be defined as a set of traits that characterize the embryos of a group of phylogenetically related species. Many times, such common traits are also evident in adults. In the most traditional sense, the traits are morphological traits and the groups of species are animal phyla (e.g. arthropods or chordates). However, the classification of the different clades of the phylogenetic tree in ranked categories such as phyla, superphyla, classes, etc. is becoming controversial. Indeed, specific body plans have been proposed for insects (Sander 1976) and crustaceans (Deutsch and Mouchel-Vielh 2003), that are both included within the arthropod phylum.

Tightly related with the concept of body plan is the concept of phylotypic stage, that is the developmental stage when the similarity between embryos that share a particular body plan reaches a maximum (Figure 1.10). Interestingly, this phylotypic stage or period always seems to happen towards the middle of the developmental process (Sander 1976, Duboule 1994, Williams 1994). Indeed, the general trend is that embryos of different species are more different at the beginning, then they converge towards a similar shape during the phylotypic stage and finally diverge again to generate varied adult forms. This phenomenon receives the name of hourglass model. Remarkably, this hourglass tendency in morphological similarity have been shown to be mirrored at the level of gene expression between different species of insects (Kalinka et al. 2010), vertebrates (Domazet-Lošo and Tautz 2010) and nematodes (Levin et al. 2012). This reinforces the intricate relationship between morphology and gene expression that we have discussed extensively. Interestingly, the

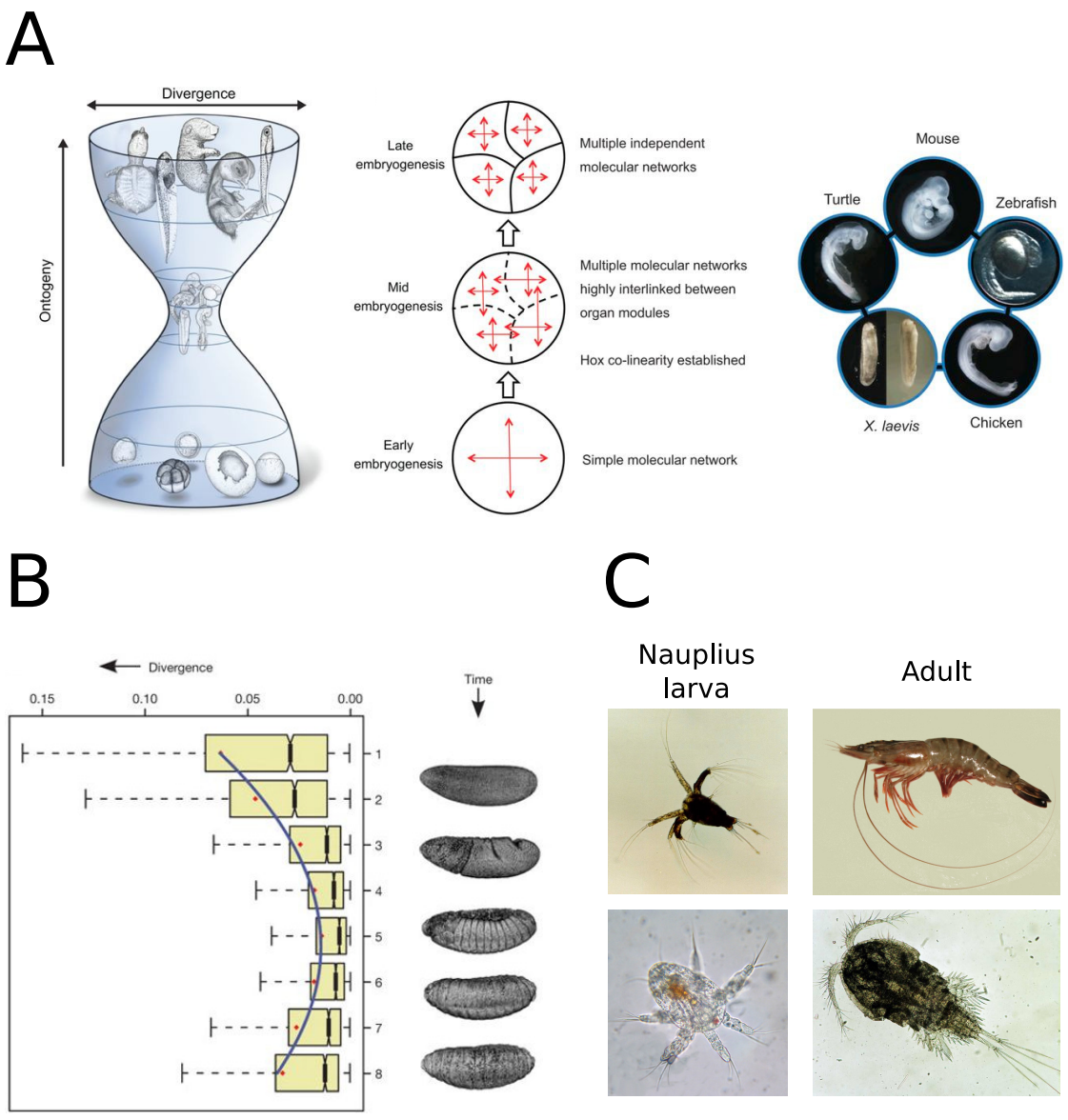


Figure 1.10: The hourglass pattern is found in several phyla. Both transcriptionally and morphologically, there is a period during mid development where the similarities between embryos of the same phyla reaches a maximum. (A) Hourglass model in vertebrates (Irie and Kuratani 2014, reprinted with permission from The Company of Biologists). (B) The hourglass also holds true for different species of *Drosophila* (from Kalinka et al. 2010, reprinted with permission from Springer Nature). (C) Also in crustaceans, the nauplius larvae are morphologically very similar between different species that are quite dissimilar in their adult forms like prawns (*Penaeus monodon*, top two images, public domain), and copepods (bottom images, the nauplius larva is work from Lithium57 distributed under the CC BY-SA 3.0 license, and the adult copepod image is public domain).

pattern is reversed when comparing the transcriptome of organisms belonging to different phyla: they are more divergent during the mid development (Levin et al. 2016). However, less is known about the circuitry of the GRN responsible for establishing those common transcriptional programs in different embryos. A comparative study between zebrafish and medaka, two teleost species that diverged more than 100 million years ago, revealed the presence of more than 700 conserved putative enhancers acting at the vertebrate phylotypic period in both species (Tena et al. 2014).

This number is probably an underestimation, since homologous enhancers are often undetected by regular sequence alignment procedures. In any case, the study revealed a complex and conserved network connectivity behind the morphological similarity of both species.

Indeed, it seems plausible to think that the stability of actual body plans is due to the extreme conservation of the GRN circuits acting during the different phylotypic stages. However, the question remains of why those circuits are more conserved than others acting earlier or later in development. Perhaps the most convincing hypothesis relies on the level of modularity of the circuits of the GRN (Raff 1996). The phylotypic period often takes place in a small embryo with a moderate amount of cells but, in contrast to earlier embryos, these cells are beginning to differentiate and take important decisions. It has been proposed that the different GRN circuits patterning this early embryo, which include diffusible signaling molecules, are highly interdependent with one another. In other words, there is little modularity at this stage of the developmental process and tweaking a single component may have unpredictable consequences throughout the whole embryo (Duboule 1994; Galis and Metz 2001). In contrast, later in development, GRN circuitry is much more modular. Therefore, changes to more peripheral circuits patterning structures such as limbs, for example, might be more easily tolerated since they will not have consequences elsewhere. This seems a plausible way to explain how body plans are stable, although further testing is needed. In contrast, it does not answer how those body plans originated in the first place. In the following section, we will focus on what is known about the evolution of the body plan of chordates in general and vertebrates in particular.

1.5

The evolution of the chordate and the vertebrate body plan

The evolutionary history behind the origin of vertebrates is still full of missing pieces, despite the fact that the evolutionary history of this group is also our history as human beings. This reflects how difficult the question was in the first place, and that is so because most vertebrates display a series of characteristics that set them well apart from the rest of the animals (Gee 2018). They include the evolution of a totally new head with a complex central nervous system and sensory organs, a bony endoskeleton, paired appendages, etc. One of the main difficulties is that there are few examples of close relatives of vertebrates (either extant or extinct) that are helpful in the endeavor of tracing the historical and molecular origin of those vertebrate novelties.

In order to clarify this we will first introduce the position of vertebrates in the phylogeny of animals (Simakov et al. 2015, Figure 1.11A). Vertebrates are included inside the big group bilateria, in contrast to non bilaterian groups of animals (e.g. sponges, ctenophores and cnidarians). Inside bilateria there is a distinction between deuterostomes and protostomes, and vertebrates are deuterostomes. This distinction used to rely in the fate of the embryonary blastopore. The blastopore is the cavity that is formed during gastrulation to give raise to the primitive gut. Protostomes were those animals where the blastopore ends up becoming the mouth and deuterostomes those where the blastopore ends up becoming the anus, an the mouth forms secondarily. Molecular phylogenetics (that infers relationships between animals based on their DNA sequence) confirmed this classification with special fortune in the deuterostome lineage. However, the situation in protostomes is much more plastic (Martín-Durán et al. 2012,Martín-Durán et al. 2016).

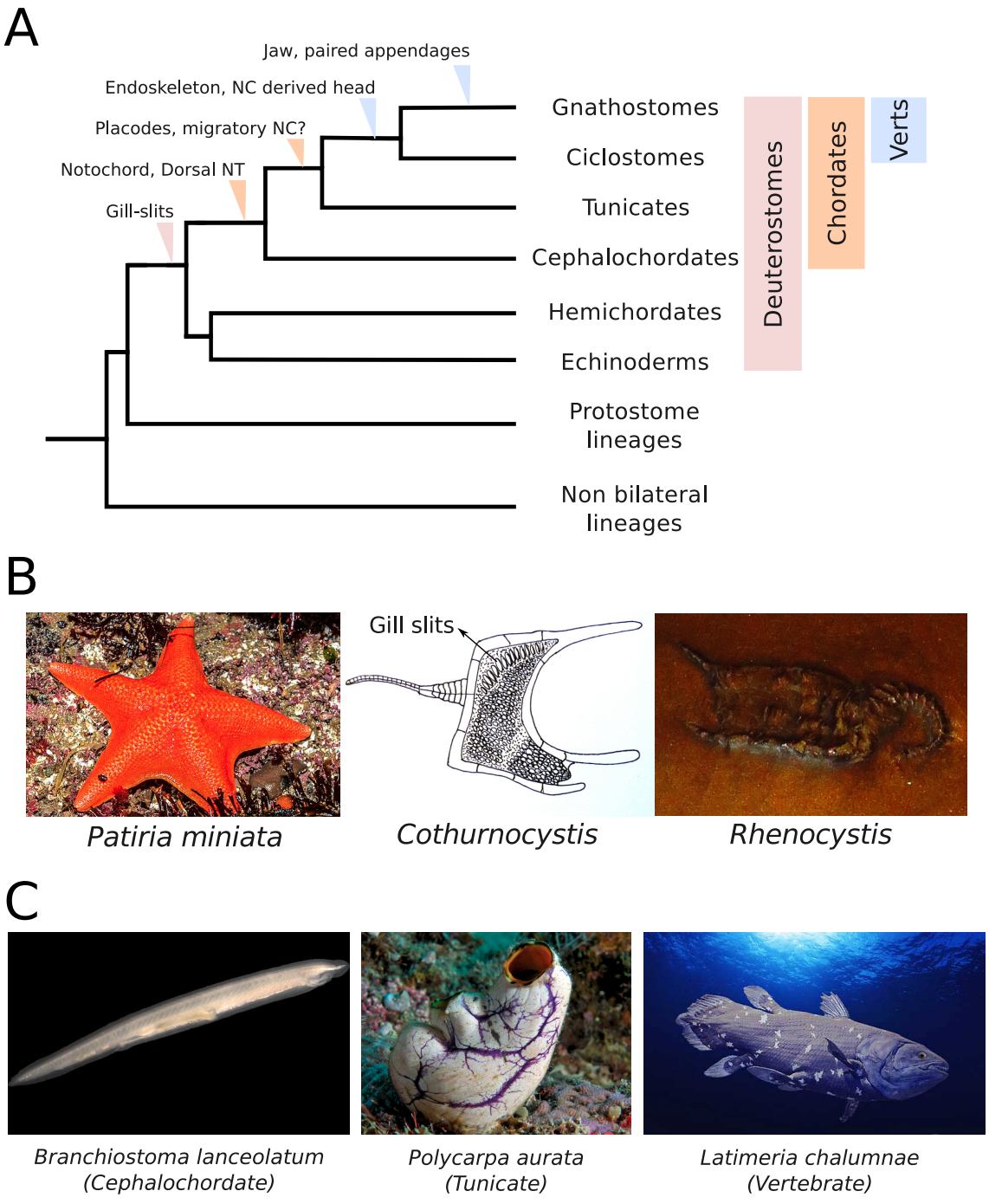


Figure 1.11: There are few close relatives to vertebrates. (A) Phylogenetic position of vertebrates in the tree of life. The appearance of some vertebrate defining traits are highlighted. (B) Echinoderms are deuterostomes, and although extant species display highly divergent morphologies like the asteroid *Patiria miniata*, there are echinoderm fossils bilaterally simetrical and even with gill slits. The *Patiria miniata* image is by Jerry Kirkhat, distributed under the CC BY 2.0 license. The *Cothurnocystis* drawing is by Haplochromis and the *Rhenocystis* image is by Ghedoghedo and both are distributed under the CC BY-SA 3.0 license. (C) Our closest relatives are the non vertebrate chordates like the cephalochordates and the tunicates. The morphology of tunicates, specially the morphology of ascidians, is really divergent despite the fact that they are phylogenetically closer than cephalochordates. The *Branchiostoma* image is by Hans Hillewaert and is distributed under the CC BY-SA 4.0 license. The *Polycarpa* image is by Nick Hobgood and is distributed under the CC BY-SA 3.0 license. The coelacanth image is by Zoo Firma and is also distributed under the CC BY-SA 3.0 license.

Current deuterostomes comprise two groups: ambulacraria and chordates. Ambulacrarians include echinoderms such as sea stars, sea urchins and sea cucumbers. They also include hemichordates like some species of acorn worms (enteropneusts) and pterobranchs (Simakov et al. 2015). Conversely, vertebrates belong to the chordates group together with cephalochordates and tunicates. Importantly, the deuterostome ancestor was probably a filter feeder organism that displayed a feature that is only found within this clade, the pharyngeal slits (reviewed in Lowe et al. 2015). Pharyngeal slits are ventral overtures that communicate the digestive system with the outside, and examples of them have been found within all major deuterostome groups. They are visible in the adult forms of extant filter feeder organisms like cephalochordates, tunicates and hemichordates. Homologous structures are also observable during the embryonic development of vertebrates although many times they never open, and they are called pharyngeal arches instead. Remarkably, vertebrate pharyngeal arches contain the precursors of many of the elements of the newly evolved head of vertebrates. Finally, no pharyngeal slits have been found in any extant echinoderm. However, there are evidences in the fossil record of extinct echinoderms with pharyngeal slits (Dominguez, Jacobson, and Jefferies 2002, Figure 1.11B). Interestingly, the patterning of this structure partially relies on the genes *nkx2.1, nkx2.2, foxA* and *pax1.9*, that are found together in the same chromosome in many sequenced deuterostomes of different groups (including some echinoderms). Therefore this configuration was likely present in the deuterostome ancestor (Simakov et al. 2015).

1.5.1

The chordate body plan and its novelties

Chordates comprise three major groups: cephalochordates, tunicates and vertebrates. Interestingly, molecular phylogenies reveal that cephalochordates conform the earliest divergent branch and tunicates and vertebrates are more closely related to each other (Delsuc et al. 2006). That was surprising at first since cephalochordate and vertebrate morphologies seem much more similar, specially when compared to adult forms of most tunicates like ascidians of larvaceans (Figure 1.11C). Indeed, ascidians were even classified as mollusks for a number of years. This probably reflects a rapid independent evolution of the tunicate lineage from the last common ancestor of tunicates and vertebrates. This ancestor was likely morphologically closer to a vertebrate or even to a cephalochordate.

The most defining characteristic that is shared by all the members of this group is the notochord (Figure 1.12A). The notochord is a stiff rod shaped structure that extends through the middle of the antero-posterior (AP) axis of chordate embryos. It appears shortly after gastrulation, which is the process by which the three germinal layers present in chordates (ectoderm, mesoderm and endoderm) are specified (reviewed in Stemple 2005). This process is highly variable between different chordate species, but generally consists in the internalization of cells that will acquire mesodermal and endodermal fates beneath those that will become part of the ectoderm. The notochord is formed by a subpopulation of those mesodermal cells that migrate towards the central part of this intermediate layer. From this privileged spot, the notochord is crucial in the proper specification of nearby cell populations, apart from providing structural support to the developing embryo. Particularly studied is the role of the gradient of *Shh* signaling released by the notochord in the dorso-ventral patterning of the ectodermal cells situated above (better said, dorsally), that

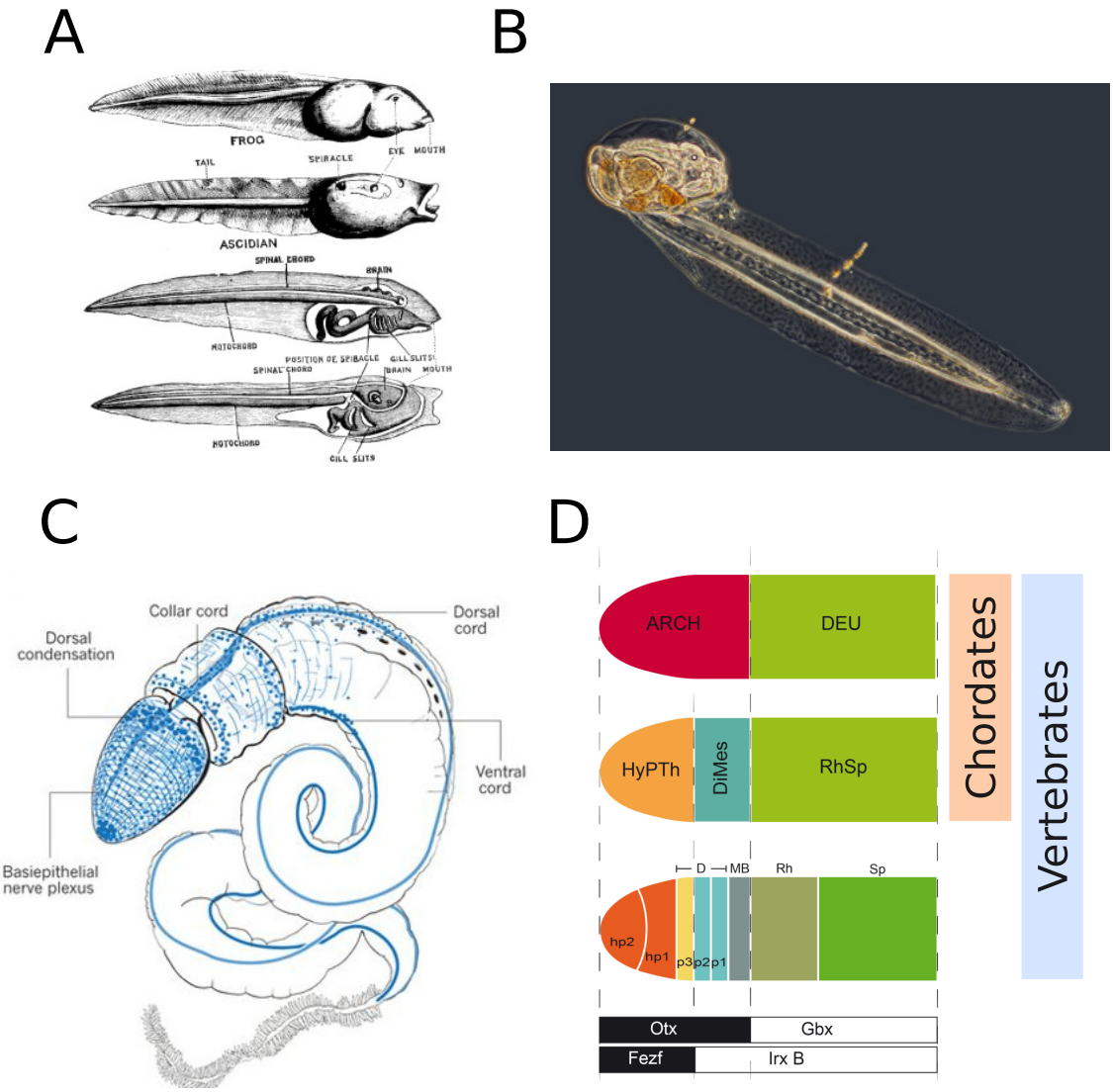


Figure 1.12: The notochord and the dorsal hollow neural tube are two defining traits of the chordate lineage. (A) Classic drawing by Lankester in 1891 showing the similarities between ascidian and anfibian tadpoles, with shared morphological traits like the notochord. (B) Larvaceans like *Oikopleura dioica* conserve the notochord throughout their adult life. Image by Proyecto Agua, distributed under the BY-NC-SA 2.0 license. (C) Drawing of the complex nervous system of the hemichordate *Saccoglossus kowalewskii*, including the ventral and the dorsal nerve chords and the small but hollow collar chord. From Lowe et al. 2015, reprinted with permission from Springer Nature. (D) Conserved regionalization of the CNS in prosomeres between amphioxus and vertebrates, with the HyPTh/DiMes and DiMes/RhSp boundaries delimited by the expression pattern of the genes *Fezf/Irx* and *Otx/Gbx* respectively (Albuixech-Crespo et al. 2017, CC BY 4.0 license).

will become the neural tube. Furthermore, other gradients of morphogens originated in the notochord are also required for the patterning of endodermal tissues and in establishing the left-right assymetry. Interestingly, cephalochordates like amphioxus and tunicates like the larvacean *Oikopleura dioica* conserve the notochord throughout their entire adult life (Escriva 2018, Bassham and Postlethwait 2000, Figure 1.12B). In contrast, in vertebrates and ascidians tadpoles is only a transient structure yet indispensable for the proper progression of development. Interestingly, a conserved expression of the transcription factor *Brachiury* is critical both for the gastrulation and for the proper notochord formation across chordates.

Another novel characteristic of the chordates is precisely the presence of a dorsal hollow neural tube formed by the ectodermal cells located over the notochord. This neural tube will eventually give rise to the adult central nervous system including the brain and the spinal cord and develops thanks to a conserved mechanism called neurulation that turns a flat neural epithelium (the neural plate) into a closed hollow tube (reviewed in Greene et al. 2017). The process roughly consists in the bending of the neural plate in such a way that both left and right lateral edges of the neural ectoderm fuse together forming a cylinder. Both the dorsal hollow tube and the neurulation process are shared by the three chordate lineages (Hudson 2016, Albuixech-Crespo et al. 2017). Importantly, no bona fide hollow neural tube can be found outside chordates (Martín-Durán et al. 2018). In protostomes with well developed central nervous systems the nerve cord is ventral and solid (and the gut dorsal). Besides, echinoderms present a very particular organization with the neural tissue mostly located beneath the skin organized in radial nerves that converge in the central nerve ring (Clark et al. 2019). Strikingly, hemichordates display two solid nerve cords along the trunk, one dorsal and one ventral (Figure 1.12C). However, no clear homology can be established between neither of the two and the neural tube of chordates although there is some debate about the collar cord of enteropneusts (Lowe et al. 2015). Briefly, enteropneusts are divided in proboscis, collar and trunk. The collar cord, that is the continuation of the solid dorsal nerve cord of the trunk in the smaller collar segment, is a hollow tube. In addition, some genes expressed in the collar cord are related with those expressed in the neural plate of chordates (Miyamoto and Wada 2013). Nevertheless, the homology between the collar cord and the neural tube of chordates is still controversial.

An important feature of the chordate neural tube is a conserved compartmentalization in domains along the AP and dorso-ventral (DV) axes (Albuixech-Crespo et al. 2017, Figure 1.12D). Broadly speaking we can differentiate 4 regions along the AP axis of chordates. From anterior to posterior we encounter the hypothalamo-prethalamic primordium (HyPTh), the diencephalic-mesencephalic primordium (DiMes), the hindbrain and the spinal cord. The limits between those regions are established thanks to the precise expression of several developmental genes. For instance, the frontier between the HyPTh and the DiMes is marked by the abutting patterns of expression of *Fezf* and *Irx* (anterior and posterior to the boundary respectively). Meanwhile, the boundary between the DiMes and the hindbrain can be equivalently traced using the patterns of expression of *Otx* and *Gbx*. In addition, the Hox genes are critical for the specification of the different embryonary segments of the hindbrain called rhombomeres. The former arrangement is common both for amphioxus and vertebrates, and a simplified version can be traced in ascidians tadpoles (reviewed in Hudson 2016). However, important elaborations over this ground plan characterize the vertebrate neural tube. Perhaps the most spectacular ones are the extensive elaboration of the telencephalon in the dorsal part of the HyPTh (Sestak and Domazet-Loso 2015) and the appearance of the eleven pairs of cranial nerves (Schlosser, Patthey, and Shimeld 2014), features that are conserved among all vertebrates. The cranial nerves are among other things responsible for the innervation of some key vertebrate structures of the head such as the facial muscles. Intriguingly, some equivalences can be drawn between the patterns of expression displayed in the neural tube of chordates, the ectoderm of the hemichordate *Saccoglossus kowalewskii* (Pani et al. 2012) and even the ventral nerve cord of *Drosophila melanogaster* (Irimia et al. 2010). This might speak for a very ancient origin of an ectodermal patterning system that was then differentially deployed and elaborated in the different lineages.

Finally, both cephalochordates and vertebrates present another unique characteristic that is

the segmentation of a specific mesodermal population of the trunk in somites (reviewed in Brent and Tabin 2002). This specific population is the paraxial mesoderm, the fraction of mesodermal cells that lies immediately adjacent to the notochord. Somites are paired and rounded and hollow aggregates of mesodermal cells that in vertebrates are later specified in different lineages: the dermomyotome and the sclerotome. The dermomyotome give rise among other things to the skin and the muscles of the trunk while the sclerotome give rise to the bones and cartilages of the vertebrae and the ribs. Interestingly, although cephalochordates lack vertebrae and ribs, in the somites of amphioxus there seems to be a cell population equivalent to vertebrate sclerotome that produces collagen (Mansfield et al. 2015). However, there are also important differences. Strikingly, the somitogenesis (the process by which the paraxial mesoderm is segmented in somites) proceeds in opposite directions in amphioxus and in vertebrates (caudal to rostral vs rostral to caudal respectively, Beaster-Jones et al. 2008). In addition, vertebrate somites are highly individualized depending on their relative position in the AP axis (Carapuço et al. 2005). This is critical for the development of the different types of vertebrae and the proper number of ribs. As we will explore in the last chapter of the introduction, this is achieved thanks to the collinear expression of the Hox family of TFs. However, this individualization seems to be lacking in amphioxus (with the exception of the most anterior pair of somites) and Hox genes are not expressed in the paraxial mesoderm (Pascual-Anaya et al. 2012).

1.5.2

The vertebrate body plan and its novelties

We have described so far a number of defining characteristics of vertebrates that are shared by other chordates, but vertebrates also display a complete set of unique innovations that set them well apart from cephalochordates and tunicates. Perhaps one of the most obvious is the presence of an endoskeleton made of cartilages and bones thanks to the appearance of the genetic program of chondrocytes and osteoblasts respectively. It has been proposed that this genetic program might be related to the program generating the cartilage-like based structures found around the cirri and the gill bars of amphioxus (Jandzik et al. 2015). However, this hypothesis needs to be further explored. In addition to that, we will briefly discuss two more vertebrate novelties: the appearance and elaboration of the vertebrate head and the paired appendages. For the sake of not extending too long we will not be covering other important innovations such as the appearance of adaptive immune cells, the highly developed renal system or the hypophysis and its underlying complex hormonal signals (Gee 2018).

The appearance of the vertebrate head seem to be highly dependent on the appearance of a new cell population, the neural crest (reviewed in Medeiros 2013, Figure 1.13A). Neural crest cells originate at the end of the neurulation from the most lateral portion of the neural plate. After the closure of the neural tube, these cells delaminate and acquire a great migratory potential (Figure 1.13B). Cranial neural crest cells, that are derived from the portion of the neural tube that corresponds to the paired rhombomeres of the hindbrain, migrate towards the pharyngeal slits (or perhaps arches) and end up forming different head structures including the bones of the skull. Intriguingly, the origin of the neural crest seem to coincide with the origin of the cranial sensory placodes (reviewed in Patthey, Schlosser, and Shimeld 2014). These sensory placodes are paired structures that give rise among other things to neuronal cells of the ear, the olfactory system and

the lens of the eye. They originate from thickenings of the non neural ectoderm that is adjacent to the neural plate (i.e. the preplacodal domain). Interestingly, placodal cells are also migratory and despite being originally non neural they can acquire neural fates to give rise to the connections between the sensory organs and the central nervous system. Both the neural crest and the cranial placodes are often considered a true novelty of vertebrates, but some rudiments can be traced in tunicates. On one hand, some cell populations around the neural tube of ascidians are also migratory and give rise to some pigmented and sensory cells (Jeffery, Strickler, and Yamamoto 2004). On the other, some thickenings of the ascidian epidermis express genes that are reminiscent of those expressed in vertebrate cranial placodes such as the olfactory placode (Abitua et al. 2015). Interestingly, this is one of the few aspects in which tunicates and vertebrates seem more alike than vertebrates and cephalochordates. Finally, apart from the tissues derived from the neural crest and the cranial placodes, the head contains an important number of muscles that allow vertebrates to perform disparate tasks such as moving the eyeballs, chewing or changing the facial expression. Strikingly, many of those muscles come from a mesodermal primordium called cardiopharyngeal field (CPF) that give rise both to those head muscles and to the muscles of the heart (reviewed in Diogo et al. 2015). Interestingly, the origin of this primordium can also be traced back at least to the common ancestor of vertebrates and tunicates (Figure 1.13C. Surprisingly, the ascidian homologous to the vertebrate CPF give rise both to the muscles of the heart and to the muscles of the atrial siphon (Stolfi et al. 2010), the structure used by these organisms to expel the water excess once the nutrients are filtered.

Next we will switch to the evolution of paired appendages (such as limbs and fins) that are critical for vertebrate locomotion among other uses (reviewed in Freitas, Gómez-Skarmeta, and Rodrigues 2014). Technically speaking, paired appendages are not a vertebrate novelty but a novelty of gnathostomes or jawed vertebrates, since agnathes lack these structures (e.g. lampreys and hagfishes). Setting aside tetrapods, which is the group that comprises the terrestrial vertebrates and aquatic mammals, the adult form of the vertebrate paired appendages are the pectoral and the pelvic pairs of fins. Apart from those paired fins, gnathostomes also display unpaired fins like the dorsal fin, the anal fin and the caudal fin. Since lampreys and hagfishes also display those unpaired fins, they are considered to be more ancient. Interestingly, both kinds of fins are composed of similar cell populations: a finfold derived from the epidermis and an endoskeletal part with mesodermal origin. It is possible to trace the origin of the finfold to the last common ancestor of chordates, since amphioxus display a dorsal finfold that covers the entire dorsal midline. In contrast, in vertebrates this finfold is restricted to the three discrete midline fins: dorsal, anal and caudal. In addition, the dorsal finfold of amphioxus lack the endoskeletal counterpart. In fact, the endoskeletal part of the paired fins is derived both from the lateral plate mesoderm and from some cells of the myotome, which does not exist in amphioxus (Onimaru et al. 2011). Interestingly, this lateral plate mesoderm is present in lampreys, even though lampreys do not form neither pectoral nor pelvic fins. In the fin to limb transition, which led to the origin of terrestrial tetrapods, the finfold was greatly reduced in favor of the endoskeletal part (Freitas et al. 2012).

Importantly, very similar GRNs operate in the endoskeletal fraction of both paired and unpaired fins (Freitas, Zhang, and Cohn 2006). For instance, the outgrowth of the fin buds is governed by a gradient of *Fgf8* that have its source in the distal most part of the structure: the Apical Ectodermal Ridge (AER). Besides, a gradient of *Shh* coming from the posterior most part of the bud (Zone of Polarizing Activity or ZPA) is critical to properly establish the AP polarity of the developing appendage. In addition, as we will explore in the last section, the expression of different Hox genes

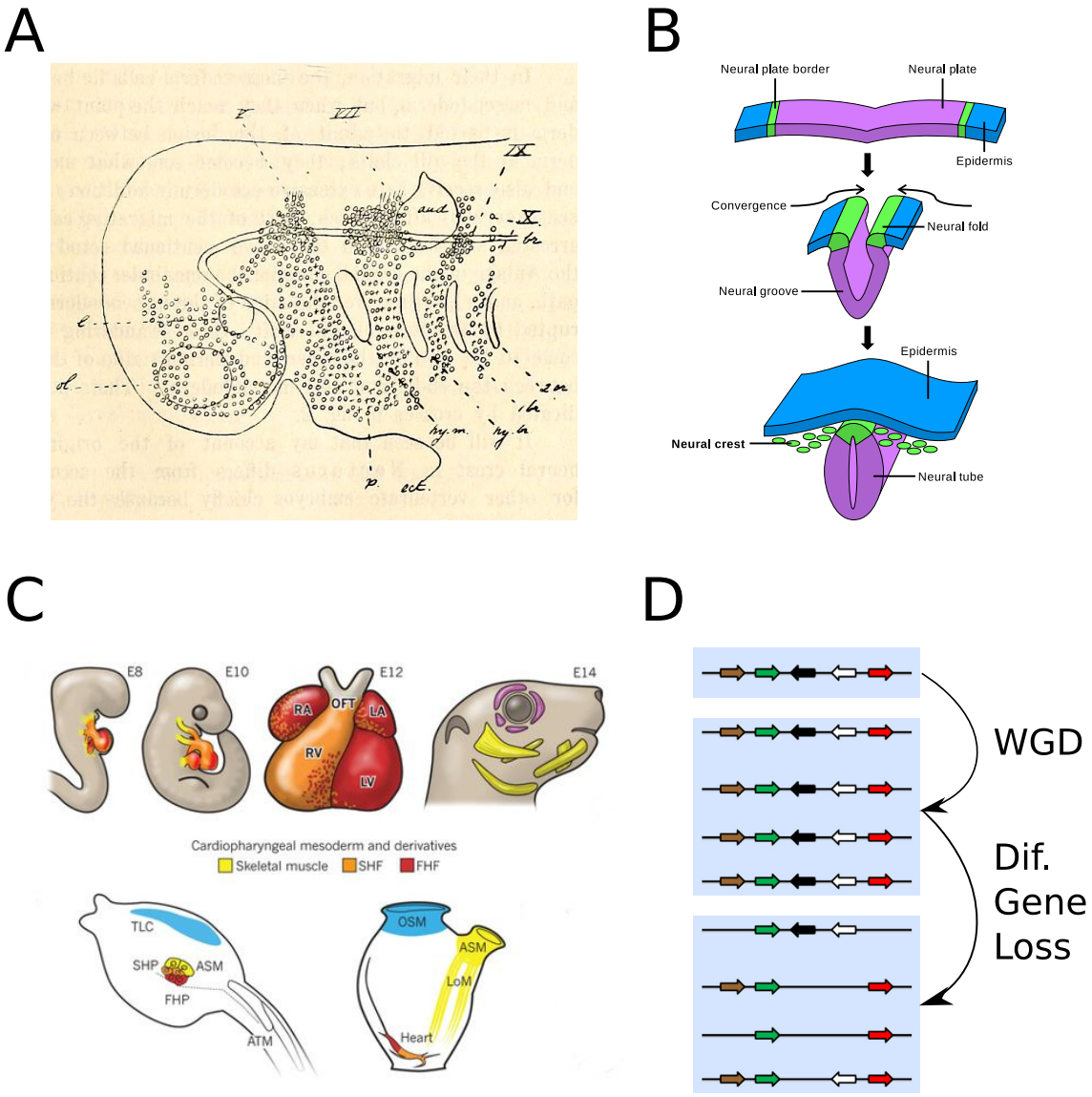


Figure 1.13: The evolution of a new head in vertebrates and the WGDs. (A) Classic drawing from Julia Platt's article first describing how migrating ectodermal neural crest cells were the ones responsible for the formation of cranial bones and cartilages in vertebrates (*Ontogenetische Differenzirung des Ektoderms in Necturus*). (B) Neurulation process highlighting the positioning and migration of neural crest cells (public domain cartoon). (C) Muscles from the heart and from the head have a common origin in the cardiopharyngeal field cell population (CPF). This CPF seem to be a feature conserved also in tunicates, and the derivatives of this population give rise also to the heart muscles and to the muscles of the atrial syphon in ascidians (Diogo et al. 2015, reprinted with permission from Springer Nature). (D) Schematic representation of the partial redundancy generated by the extra gene copies that are generated after a WGD event.

is essential in defining distinct cell fates along the proximo distal axis. This is better studied in tetrapods where the limbs are divided in stylopod, zeugopod and autopod (reviewed in Tanaka 2016). Those will be equivalent to the arm, the forearm and the hand of the human forelimb and to the thigh, the calf and the foot of the human hindlimb. The similarities between the genes patterning the different type of fins led to the hypothesis that the GRN that functions in the unpaired fins was co-opted to the lateral plate mesoderm, propitiating the origin of the paired fins (Freitas, Gómez-Skarmeta, and Rodrigues 2014). This hypothesis is now strongly supported by

recent experiments regarding the ZRS in teleost fishes, which is the enhancer driving the expression of *Shh* to the ZPA. In zebrafish and medaka, the ZRS enhancer drive the expression of *Shh* to the posterior end of all the fin buds, both paired and unpaired (Letelier et al. 2018a). Furthermore, removing the ZRS enhancer in medaka caused both developmental defects in the pectoral fins and the complete absence of the dorsal fin. Then, this tight linkage between paired and unpaired fins reflects that they share a rather similar transcriptional program.

Lastly, we will comment upon another important peculiarity of vertebrates that lies inside their genomes. It seems pretty clear that in the transition from the last common ancestor of chordates to the last common ancestor of gnathostomes two rounds of whole genome duplication (WGD) occurred (Dehal and Boore 2005). Therefore, in principle, four potential copies of each gene found in the genome of non vertebrate chordates appeared in the gnathostome genomes (Figure 1.13D). Importantly, it has been for long argued that the events of WGD potentially boost the appearance of novelties because of the rapid redundancy that is generated (Ohno 1970). This cause that mutations that otherwise would have been lethal might be compensated by the multiple copies that originated (called paralog genes), allowing higher rates of neofunctionalization. Indeed, important novelties seem to rapidly originate at the root of vertebrates along with the events of WGDs. It is important to note that teleost fishes underwent an extra round of WGD (Amores 1998) and that lampreys also present three WGDs with respect to the chordate ancestor (Pascual-Anaya et al. 2018). However, it is still a matter of debate if one or two of the three WGDs experienced by the lamprey lineage are shared with the rest of gnathostomes.

To sum up, a wealth of novelties were incorporated during the evolution of both the chordate and the vertebrate body plans. All of them are accompanied by the expression of specific genes that determine the fate of the cell populations that will construct them. Therefore, changes at the regulatory level were surely needed. Importantly, the primordia of many of such novelties of both the chordate and the vertebrate body plan are specified or patterned during the vertebrate phylotypic stage (i.e. the somites, the pharyngeal arches, the limb buds, the neural crest and the placodes). In the last section of the introduction we will explore how the Hox genes participate in the patterning of some of those structures; particularly the somites, the hindbrain, the cranial neural crest cells and the paired appendages. In addition to that, Hox are extremely interesting because of the very special role of the 3D architecture of the vertebrate loci in regulating the expression of these genes.

1.6

Ancient and novel roles of Hox genes in the building of the vertebrate body plan

Hox genes constitute an ancient group of TFs belonging to the homeobox superclass (Holland 2013). Homeoboxes in general and Hox proteins in particular are characterized by the presence of the homeodomain, which is a proteinic domain that mediates the binding of these proteins to the DNA. The homeodomain comprises three alpha-helices and an unstructured N-terminal arm, and by itself it binds the DNA weakly and without much specificity. In the case of the Hox genes, this is partially overcome due to the interaction with a series of cofactor proteins that include other

homeoboxes such as the ones encoded by the also ancient *hth/Meis* and *exd/Pbx* gene families (for review see Merabet and Mann 2016).

Hox genes can be classified in four types: anterior, Hox3, central and posterior. Indeed, their names reflect the unusual way they are often found in the chromosomes: together forming clusters and consistently ordered from anterior to posterior (Figure 1.14A). They were first identified in *Drosophila* when mapping homeotic mutations (i.e. those that caused the transformation of one body segment into another). An example of an homeotic mutant is the four winged fly that we explored earlier, caused by the loss of expression of *Ubx* (that is in fact a central Hox gene) in the third thoracic segment (Lewis 1978). One of the two main reasons that made the study of Hox genes extremely appealing was precisely the fact that simple perturbations of their expression patterns was sufficient to produce drastic changes in the organization of the body plan. That reflects that Hox genes need to be high enough in the GRN hierarchies and influence the expression of an important number of effectors downstream. The other reason was the puzzling fact that the position of these genes within their cluster mirror their patterns of expression in the *Drosophila* segments along the AP axis (reviewed in McGinnis and Krumlauf 1992). In other words, Hox genes located at the anterior end of the cluster are expressed in anterior segments, central genes in central segments and posterior genes in posterior segments. That was remarkable, as it was the finding that Hox homologs display similar clustering and expression dynamics also in vertebrates (Duboule and Dollé 1989, Figure 1.14A). Indeed, homeotic transformations also happen in mice when Hox genes are mutated or their expression is altered (Carapuço et al. 2005). More recently, similar expression dynamics have been found in several spiralians such as the annelid *Capitella teleta* (Fröbius, Matus, and Seaver 2008, Figure 1.14B) and the mollusk *Acanthochitona crinita* (Fritsch et al. 2015).

Finding those striking similarities between groups as divergent as deuterostomes, ecdysozoans and spiralians seem to speak for a common origin of the regulatory mechanism that links genomic order with ordered expression patterns along the AP axis. This phenomenon was termed collinearity. Importantly, the ancestral domains of collinear expression of Hox genes are thought to be ectodermal and neural, since these domains are found in a wider range of species. However, both in *Drosophila* and in vertebrates, Hox genes collinear expression is also important in the AP patterning of mesodermal structures. For instance, the larval *Drosophila* circulatory system is formed by the anterior aorta, the posterior aorta and the heart. Those structures extend axially from the first thoracic segment (T1) to the seventh abdominal segment (A7) and the boundaries between them are demarcated by abutting expression of *Antp, Ubx* and *AbdA* respectively (Lo et al. 2002). This is an example of how the Hox collinear patterning is flexible and can be redeployed in the development of new structures. In fact, vertebrates are perhaps the textbook example of redeployment of Hox genes in new territories. This has been traditionally related to the flexibility achieved after the two rounds of vertebrate WGDs, which generated four copies from the original Hox cluster (named from HoxA to HoxD), all of them conserved and functional in extant vertebrates (Duboule 2007). In the case of zebrafish, due to the extra round of WGD, seven copies are conserved (Amores 1998). Now we will address which are the roles of Hox genes during vertebrate development in a number of territories where the collinear logic have been preserved and thoroughly studied. We will also highlight how in some of these territories a precise 3D folding around the Hox loci is critical.

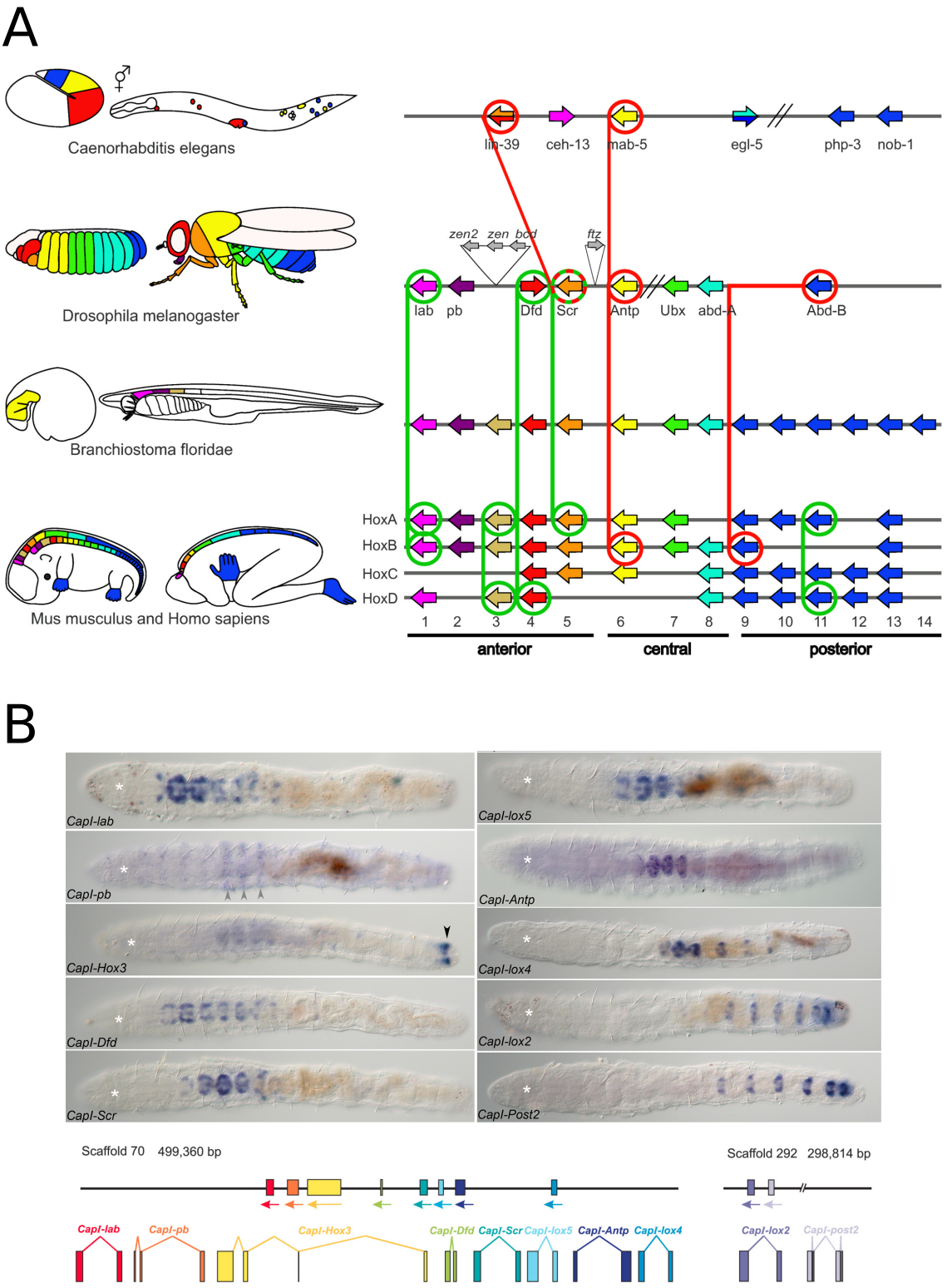


Figure 1.14: Collinearity in the expression of Hox genes is found in disparate phyla. (A) Ordered expression of Hox genes along the main body axis of disparate organisms such as *Drosophila*, amphioxus or mouse. Hox genes are often classified between anterior, central and posterior (Hueber et al. 2010, distributed under CC BY 4.0 license). (B) The recent discovery of Hox collinearity in both embryos and larvaes from the annelid *Capitella teleta* (Fröbius, Matus, and Seaver 2008 CC BY 4.0 license).

1.6.1

Hox genes in the patterning of the CNS and the neural crest

We already covered briefly how the abutting pattern of expression of different genes pattern the neural tube of chordates (Albuixech-Crespo et al. 2017), and we anticipated that the Hox genes were involved in the patterning of the eight different segments of the hindbrain called rhombomeres (r1 to r8). Indeed, anterior most Hox genes from the four clusters are expressed collinearly in the developing hindbrain of vertebrates (reviewed in Parker, Bronner, and Krumlauf 2016). First, *HoxA1/HoxB1* are expressed in a broad domain with an anterior limit located in the presumptive boundary between the rhombomeres 2 and 3 (r2/r3). These genes are turned on by a wave of the retinoic acid (RA) morphogen that extends rostrally and dorsally from the anterior most somites. Then, their expression is restricted to the r4 due to the activity of CYP26 enzymes near the rhombomere r1, that metabolize RA preventing further activation of *HoxA1/HoxB1*, and to the expression of the TF *Krox20* in the rhombomeres r3 and r5. *Krox20* suppress the expression of the *HoxA1* and *HoxB1* genes in these two rhombomeres while activating *HoxA2* and *HoxB2*. In addition, *HoxA3* and *HoxB3* genes are also activated in the r5 thanks to the coexpression of *Krox20* and *Kreisler* in this segment. Then, *HoxB4* and *HoxD4* expression start and get stabilized in the caudal most rhombomeres (r7/r8) thanks to the stronger RA concentration present in that area. Finally, more posterior Hox genes are involved in the patterning of different spinal cord motoneuronal populations (Tschopp, Christen, and Duboule 2012). For instance, different paralogs of the Hox6 and Hox10 groups define the motoneuronal cells innervating the forelimbs and the hindlimbs respectively. Importantly, perturbing the expression domains of the different Hox genes in the hindbrain lead to changes in the identity of the rhombomeres and the neuronal populations that originate there (Parker, Bronner, and Krumlauf 2016), similarly to what happens with Hox homologs in insects and the different segments of the insect body plan. This demonstrates that they are again at the top of the hierarchy of a GRN that establishes the identity of different segments along a longitudinal axis.

Strikingly, the origin of this collinear expression in the hindbrain could be arguably traced back to the deuterostome ancestor. Hox genes of the enteropneust *Saccoglossus kowalewskii* are expressed collinearly in the posterior most region of the neuroectoderm in a domain shared with *Gbx*, a gene that is expressed in the hindbrain of chordates (Lowe et al. 2003). However, it is still challenging to establish homologies between the well defined dorsal hollow neural tube of chordates and the two solid nerve cords and the nerve net of hemichordates. Much clearer is the homology between the neural tube of vertebrates and those of cephalochordates and tunicates. Indeed, collinear expression of Hox genes can be found in the amphioxus neural tube and, in remarkable similarity with vertebrates, RA signalling plays a crucial role in establishing the expression boundaries (Pascual-Anaya et al. 2012). Nevertheless, rhombomere segmentation is not evident and there seem to be no clear cross-regulation between *Krox* and Hox genes, since their expression domains do not overlap (Knight et al. 2000). Particularly, *Krox* is expressed discontinuously in the amphioxus neural tube, but in more anterior territories including the HyPTh primordium and the DiMes. This could reflect an stepwise elaboration of the hindbrain patterning, integrating first RA signalling in the chordate ancestor and later on the Krox/Hox cross-regulation. In fact, several studies performed in lampreys suggest that the vertebrate ancestor likely had a segmented hindbrain organized by a very similar GRN to that observed in model vertebrates such as zebrafish

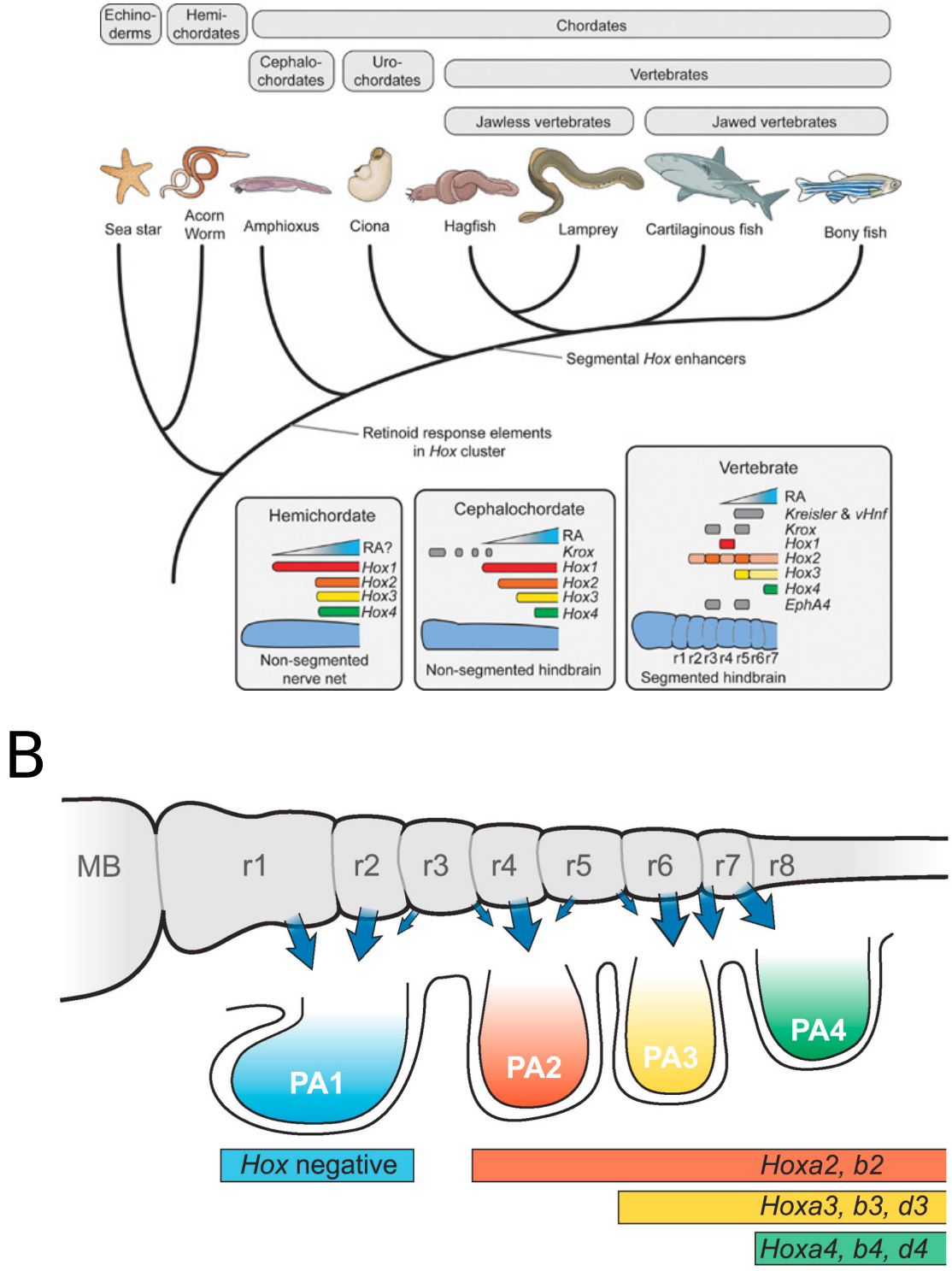


Figure 1.15: The collinear expression of Hox genes is required for the patterning of the vertebrate hindbrain and cranial neural crest cells. (A) Evolution of the GRN responsible for the hindbrain patterning. Hox genes are expressed in the CNS of hemichordates despite the lack of segmented rhombomeres, although it is unclear whether RA plays a role. In amphioxus the hindbrain is not segmented either but RA is essential for the collinear expression of Hox genes in the CNS. Finally, both in gnathostomes and agnathes, the segmented hindbrain is patterned by Hox genes in combination with the TFs *Krox* and *Kreisler* (Parker, Bronner, and Krumlauf 2016, reprinted with permission from John Wiley and Sons). (B) The Hox collinear code is also important for the proper migration of neural crest cells. For instance, the cells delaminating from the r4 and the r6 express Hox2 and Hox3 paralogs and migrate to the pharyngeal arches 2 and 3 respectively (Parker, Pushel, and Krumlauf 2018, reprinted with permission from Elsevier).

and mouse (Parker, Bronner, and Krumlauf 2014). Apart from the fact that rhombomeres are readily distinguishable in lampreys, the expression of Hox genes, *Krox* and *Kreisler* is comparable with the domains observed in jawed vertebrates. In addition, zebrafish enhancers that drive Hox expression to specific rhombomeres works equivalently in lampreys further indicating a conserved circuitry of the GRN (Figure 1.15A).

Finally, it is important to note that the regionalization of the hindbrain is critical for establishing the identity of the cranial neural crest populations (reviewed in Parker, Pushel, and Krumlauf 2018). These populations delaminate and migrate mainly from the even rhombomeres to specific pharyngeal arches in order to participate in the building of the vertebrate head (Figure 1.15B). For example, neural crest cells delaminating from the r4 (which express *HoxA1/HoxB1* genes) populate specifically the second pharyngeal arch. Intriguingly, neural crest cells migrating from the r4 stop expressing *HoxA1/HoxB1* and start to express *HoxA2/HoxB2* during the migration. In contrast, neural crest cells originated in the r6 continue to express *HoxA3* while migrating to the third pharyngeal arch. Neural crest specific enhancers are responsible for maintaining or restarting the collinear expression of Hox genes in those migrating cells. Importantly, homeotic transformations between the different pharyngeal arches have been described when Hox expression in neural crest cells is altered (Gendron-Maguire et al. 1993). This further indicates that regulated Hox expression is critical both for guiding the migration and determining the fate of these cells.

1.6.2

Hox genes in the patterning of the somites

In vertebrates, Hox genes are also expressed collinearly in the different pairs of somites that are generated along the embryonic AP axis, specifying their future fate (reviewed in Mallo 2018). Among other structures, vertebrae and ribs derive from the embryonic somites, and therefore the Hox genes expressed by a particular somite will determine the type of vertebrae that will originate and whether this vertebrae will have ribs attached or not (Figure 1.16A). In stark contrast, Hox genes are not expressed in the somites of cephalochordates (Pascual-Anaya et al. 2012) and, perhaps accordingly, most of these pairs give rise to equivalent structures with the exception of the anterior most one.

Both the generation and the specification of the somites from the seemingly disorganized presomitic mesoderm (PSM) seem to be highly overlapping processes in vertebrates (reviewed in Aulehla and Pourquié 2010). Briefly, pairs of somites start to form thanks to a cyclical process that occur in the anterior most part of the PSM: the determination front. This determination front is placed at the interface between two mesodermal domains, the anterior one dominated by RA signalling and the posterior one dominated by Wnt and Fgf signalling. RA signalling promote the differentiation of somites while Fgf and Wnt maintain the posterior PSM in an undifferentiated state. Cyclical fluctuations of gene expression involving Notch signallling and others at the interface between the Fgf and Wnt gradients allow to coordinatedly instruct evenly sized groups of presomitic precursors to differentiate into somites. Rather surprisingly, modifying the expression of Hox genes in cell populations of the PSM and not in the already formed somites is needed in order to generate homeotic transformations in mice (Carapuço et al. 2005, Figure 1.16B). This highlights the close relationship between the rhythmic production of somites and the specification of their future fate. In concordance, altering either of the three mentioned signalling pathways during somi-

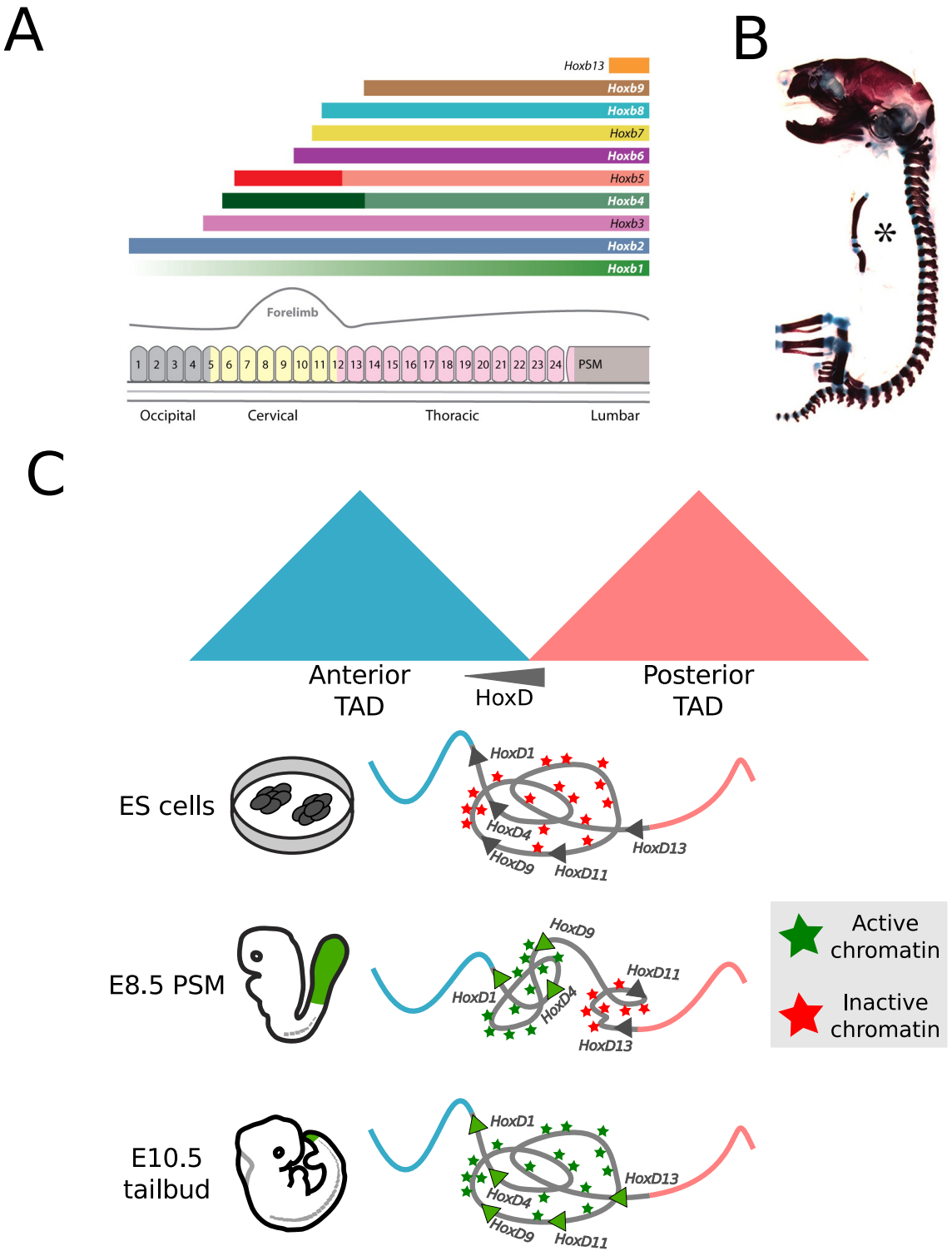


Figure 1.16: Hox genes patterning of the somites. (A) The expression of Hox genes in the PSM determine the identity of the somites and derived structures like vertebrae (Alexander, Nolte, and Krumlauf 2009, reprinted with permission from Annual Reviews). (B) Misexpression of Hox genes in the PSM are able to generate homeotic transformations like the one in the image with thoracic vertebrae transformed in lumbar through the anticipated expression of *HoxA10* in anterior territories leading to the loss of the ribs (Carapuço et al. 2005, image distributed under the CC BY 4.0 license). (C) The role of chromatin architecture ensuring the collinear expression of HoxD genes in the mouse somites. The expression is governed mainly by proximal enhancers (grey region in the cartoon). In ES cells, when the cluster is inactive, all the promoters are placed in the same 3D environment that is marked with repressive epigenetic marks. When they start to become active in the PSM, the 3D environment split in two, an active anterior one and an inactive posterior one. As development progresses, more posterior Hox genes switch to the active anterior compartment. Redrawn after the work done in Noordermeer et al. 2014.

togenesis is able to produce altered boundaries of expression of Hox genes and in turn homeotic transformations between pairs of somites. In addition, *Oct4* and *Gdf11* seem to be two critical regulators of the expression of Hox genes in the PSM, being required for the proper production of trunk and tail somites respectively (Aires et al. 2016). Indeed, a sustained expression of *Oct4* in the PSM is needed in order to originate the remarkably high amount of rib bearing vertebrae found in snakes in comparison with other vertebrates. This increase in rib bearing vertebrae can be also observed in genetically engineered mice lacking *Gdf11*.

In order to better understand how Hox genes are collinearly regulated in the PSM several seminal studies have been carried out that profile the epigenetic modifications and the 3D organization of the murine HoxD locus in those cell populations (Figure 1.16C). In mouse, the HoxD gene cluster sits precisely over a topological border that separate two big TADs often referred to as anterior and posterior respectively (the anterior and posterior TADs lying adjacent to anterior and posterior Hox genes respectively, Lonfat and Duboule 2015). This configuration would potentially allow anterior and posterior most HoxD genes to interact with distal enhancers lying in the anterior and in the posterior TAD respectively and seem to be conserved across most vertebrates (Woltering et al. 2014). However, the regulation of HoxD genes in the murine PSM seem to rely primarily in enhancers located within the cluster itself (Spitz et al. 2001). Nevertheless, that does not mean that chromatin and 3D organization does not play an indispensable role. Surprisingly, proximal contacts established by HoxD promoters revealed using 4C-seq are rather dynamic when comparing embryonic stem cells (ES cells) and PSM cells of different developmental timepoints (Noordermeer et al. 2014). In brief, in ES cells both anterior and posterior HoxD promoters interact strongly and almost evenly along the extension of the HoxD cluster. These interactions coincide with a bivalent chromatin domain populated both with H3K27me3 and H3K4me3 epigenetic marks, and HoxD genes are mostly not transcribed. In contrast, in the PSM cells there are two separated chromatin domains, one of them transcriptionally active and populated with the H3K4me3 mark and an inactive one populated with H3K27me3. From the 3D perspective the HoxD cluster is also splitted in two, with HoxD promoters present in the active part interacting primarily with the active chromatin compartment and viceversa. Concordantly with the collinear activation of HoxD genes, in PSM cells the active compartment grows from the anterior to the posterior end of the cluster during development. This in turn allows to switch on anterior HoxD genes first in those PSM cells that will give rise to the anterior somites, and more posterior HoxD genes later in those cells that will generate more posterior somites.

1.6.3 Hox genes in the patterning of the paired appendages

Epigenetics and chromatin architecture are very important for the adequate regulation of Hox genes in the neural tube (Tschopp, Christen, and Duboule 2012) and even more clearly in the somites (Noordermeer et al. 2014). However, the regulation of HoxD genes in the mouse limb is the case that perhaps exemplifies those relationships best (Figure 1.17). Reminiscent to what happens in the somites, HoxD genes are both collinearly expressed in the limb buds with anterior genes patterning the future arm (also known as stylopod) and forearm (zeugopod) structures and posterior HoxD genes instructing the making of the hand and digits (autopod). However, this time the patterning of the limb requires that the HoxD cluster is activated twice (Andrey et al. 2013).

During an early wave of gene expression, genes from *HoxD1* to *HoxD11* are collinearly activated in the proximal cell populations of the limb bud, that will give rise to the arm and the forearm. Later on, genes from *HoxD8* to *HoxD13* are also collinearly turned on in a distal cell population that will become the hand and the digits. Interestingly, it has been hypothesized that this splitted expression of the HoxD genes in two waves ensures that there is a cell population in the limb bud that does not express HoxD genes (Woltering et al. 2014). This population will in turn generate the articulated wrists and ankles found in tetrapods.

In stark contrast with the enhancers driving HoxD expression in the somites, HoxD limb enhancers are located far away from the cluster (Montavon et al. 2011). Remarkably, enhancers driving the early phase of HoxD expression are located in the anterior TAD while enhancers driving the late phase are placed in the posterior TAD. This is consistent with the fact that anterior HoxD genes, which contact preferentially genomic regions from the anterior TAD, are the ones that are expressed during the early wave in the population of proximal cells. Likewise, posterior HoxD genes that are involved in the late wave of expression in distal cells contact preferentially loci from the posterior TAD. Rather strikingly, genes that are expressed in both waves (*HoxD8* to *HoxD11*) are able to switch from contacting preferentially the anterior TAD in proximal cells to contact preferentially posterior TAD enhancers in the distal cell population (Andrey et al. 2013). An intricate array of both dynamic and stable CTCF sites seem to be responsible of both the generation of the boundary and its flexibility (Rodríguez-Carballo et al. 2017). Finally, it has been shown that Hox13 proteins are required to turn off the enhancers from the anterior TAD at the same time that it sustains the activity of the posterior TAD enhancers (Beccari et al. 2016). In the absence of both *HoxA13* and *HoxD13* genes, the anterior TAD is cannot be switched off and the posterior TAD never gets activated, leading to the loss of the wrist and the digits. Accordingly, the HoxA cluster is also located in the boundary between an anterior and a posterior TAD. This suggest an ancient origin of both the chromatin arrangement in two TADs and of the two waves of Hox expression in limbs, previous to the appearance of the two rounds of WGDs.

Accordingly, a very similar chromatin organization in two TADs is found in the HoxD and the two HoxA clusters of zebrafish (Woltering et al. 2014). However, establishing how the two coordinated waves of expression take place in fins and compare it to the mouse limbs is challenging due to the great morphological differences between both structures. In contrast to tetrapods, the fins of most ray finned fishes are composed of both two rows of endochondral bones (proximal and distal radials respectively) and a dermal finfold (Freitas, Gómez-Skarmeta, and Rodrigues 2014). That made difficult to find the zebrafish cell populations that are homologous to the proximal and distal cell populations that respond to the first and the second wave of Hox expression in mouse. However, a conserved cis-regulatory logic seem to be operating in order to activate a second wave of Hox expression (Gehrke et al. 2014). A *Shh* dependent enhancer from garfish (*Lepisosteous oculatus*) is able to drive the expression of *HoxA13* in the mouse autopod, equivalently to the endogenous one from mouse. In addition, lineage tracing experiments in zebrafish using this very same enhancer revealed that there is a cell population in ray finned fins homologous to the distal cells of the mouse limb bud (Nakamura et al. 2016). Strikingly, these cells contribute not only to a certain part of the distal radials but also generate osteoblasts of the finfold. Concordantly, the combined loss of function of *hoxa13a, hoxa13b* and *hoxd13a* in zebrafish lead to a massive finfold reduction and problems in the development of the distal radials.

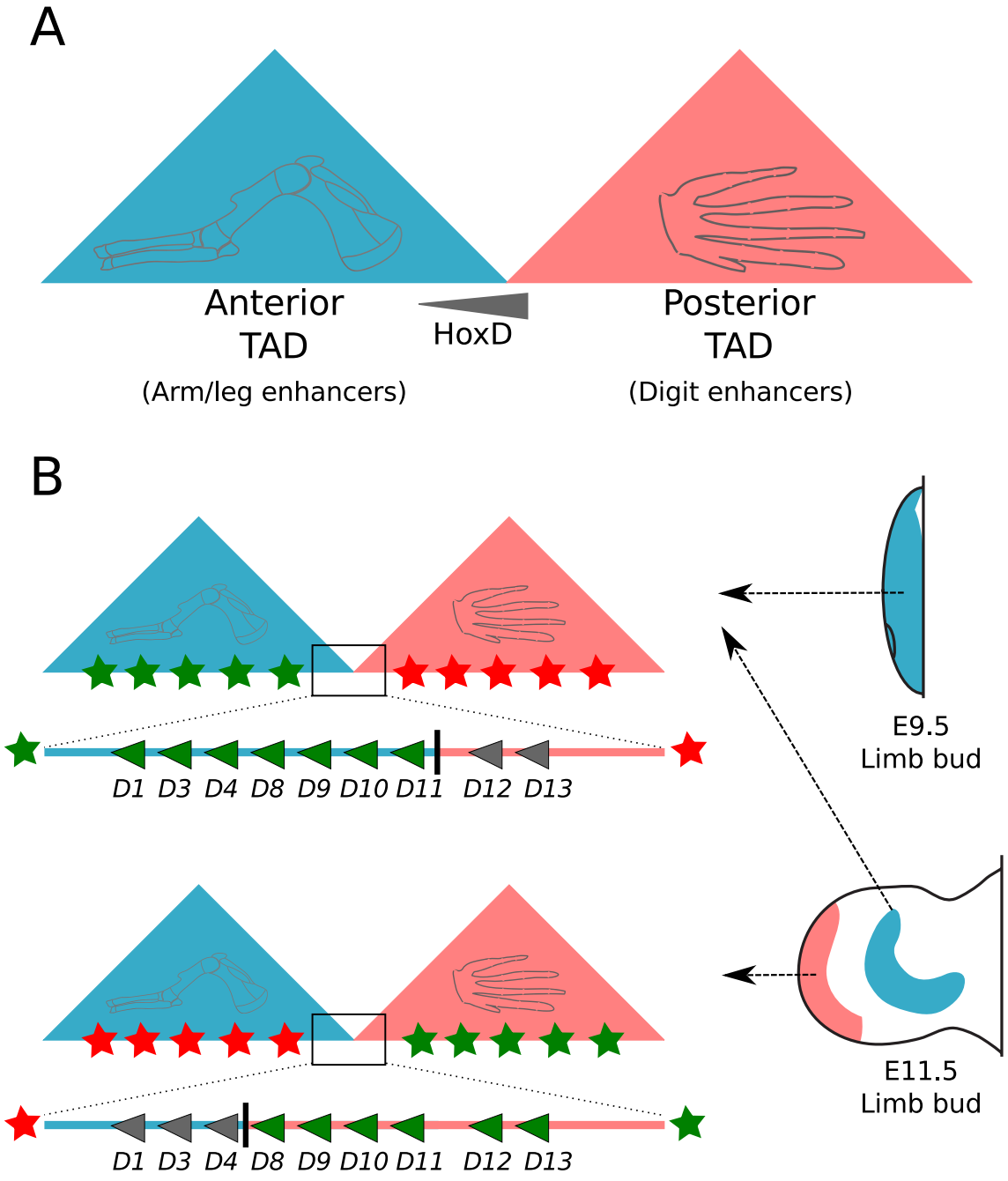


Figure 1.17: The two waves of HoxD expression in the patterning of the limbs require an specific configuration of the chromatin in two TADs with Hox promoters in the middle. During the first wave, the proximal population of cells of the limb bud expresses HoxD genes collinearly from *HoxD1* to *HoxD11* responding to enhancers located in the anterior TAD. Later, in a second wave, a distal population of cells also express collinearly the genes from *HoxD8* to *HoxD13*. The enhancers driving the expression to the distal population of the limb bud are placed in the posterior TAD. Then, intermediate HoxD genes (from *HoxD8* to *HoxD11*) need to switch from one TAD to the other to respond to both expression waves. Therefore, the location of vertebrate HoxD genes at the hinge between two TADs is critical for their proper expression in the patterning of the limbs. Redrawn after the work published in Andrey et al. 2013.

Chapter 2

Objectives

In the introduction we presented a number of precedents that motivated the research that was carried out during this thesis. In summary, a better understanding of how the non-coding genome affects the regulation of transcription is still needed in order to link changes in DNA with the appearance of morphological novelties. In this regard, the role of changes in chromatin architecture (compartments, TADs and loops) in the evolution of gene expression has been poorly explored yet despite their huge potential impact in transcription due to the rewiring of developmental GRNs. In addition, the origin of jawed vertebrates entailed the rapid appearance of a number of important morphological novelties including the neural crest or the paired appendages, coupled to two rounds of WGD. It has been already hypothesized that WGDs could have helped to relax certain constraints by allowing to tweak the expression of the different paralog genes independently, favoring the appearance of such novelties. Particularly Hox genes, which are key developmental regulators across phyla, have been retained in four copies in most vertebrates and in seven copies in some teleost fishes such as zebrafish. Interestingly, they participate in the patterning of several of the new anatomical features of vertebrates including the paired appendages. Furthermore, a rather special and dynamic chromatin architecture in two TADs at both the HoxA and the HoxD loci is critical for the regulation of these genes and the patterning of limbs and fins. How this system appeared was unknown and is tightly coupled to the origin of paired appendages. Taking this information into account, we set two main objectives:

1. To investigate when and how did the chromatin architecture found around vertebrate HoxA and HoxD clusters appear in evolution.

   In order to do that we used 4C-seq experiments coupled to computational modelling to reconstruct the chromatin architecture around the only amphioxus Hox cluster, and compared it with the one found in vertebrates.

2. To evaluate what was the impact of global changes in chromatin architecture in the regulatory innovations that appeared at the root of vertebrates.

   To achieve that we used both 4C-seq experiments and HiChIP experiments targeting different histone modifications in mouse, zebrafish and amphioxus embryos.

Chapter 3

Materials and methods

The technical part of this thesis project focused on comparative Chromosome Conformation Capture experiments performed in several flavors depending on the scope: 4C-seq (one locus to all loci), HiChIP (many to all) and HiC (all to all). Although no experimental HiC experiment is presented, some publicly available HiC datasets were reanalyzed for several purposes and therefore these analysis are also covered in this chapter. Several computational analysis of microsynteny conservation are also included and described in detail because they were key to understand the evolution of chromatin folding in combination with the C-experiments. Indeed, an special effort has been made in order to clarify extensively the computational part of the methodology and make it reproducible. For that purpose, a complementary git repository containing the code for the different analysis and usage examples is provided in *gitlab.com/rdacemel*.

In addition to that, transgenic reporter assays for enhancer detection were also performed and described. Finally, it is worth noting that we also present the result of a Whole Mount In-Situ Hybridization in amphioxus embryos that was published in Acemel et al. 2016. This experiment was performed by our collaborators in the laboratory of the Professor Hector Escrivà and the protocol is not included here. Nevertheless, it is detailed in the aforementioned publication.

3.1

4C-seq

The rationale behind C-techniques and the 4C-seq technique in particular was already presented in the pertinent section of the Introduction (1.3.1, p.15). Here we will focus on the experimental details of the protocol itself and also in the data analysis procedure.

3.1.1

4C-seq library preparation

4C-seq experiments were performed in whole embryos from four different species: zebrafish (*Danio rerio*), amphioxus (*Branchiostoma lanceolatum*), mouse (*Mus musculus*) and a marine centipede (*Strigamia maritima*). The majority of the protocol was equivalent (Figure 3.1A, based on Werken

et al. 2012 with modifications) except from the first steps before the sample fixation in Paraformaldehyde (PFA). Therefore, these first steps are described separately.

1a - Sampling and fixation of zebrafish embryos:
Synchronized zebrafish embryos were staged and sampled at either 24 hours post fertilization (24hpf) or 80% epiboly stages. 500 and 2000 embryos were pooled for each of the 24hpf and 80% epiboly experiments respectively. They were collected in 50 mL of E3 media ($NaCl$ 5mM, KCl 0.17mM, $CaCl_2$ 0.33mM, $MgSO_4$ 0.33mM, pH 6.8–6.9) supplemented with 500μL of 30 mg/ml pronase (Roche, 11459643001) in order to remove the chorions. They were incubated at 28°C during 15' until the chorions softened, and then the chorions were removed by carefully pipetting up and down with a Pasteur pipette. Dechorionated embryos were then rinsed with fresh E3 media without pronase and transferred to a 1.5mL microcentrifuge tube. The remaining E3 was carefully removed from the tube and replaced with 1mL of Ginzburg Fish Ringer buffer ($NaCl$ 111mM, KCl 3.6mM, $CaCl_2$ 2.7mM, $NaHCO_3$ 1.9mM) in order to remove the yolks. The process was helped by pipetting up and down energically with a yellow tip and incubating the sample with shaking for 5'. Cells ready to be fixated were then recovered by centrifugation at 300g for 30s and transferred to a 15mL Falcon tube containing 10mL of 2% PFA in PBS ($NaCl$ 137mM, KCl 2.7mM, Na_2HPO_4 10mM, KH_2PO_4 1.8mM). The fixation was maintained for 10' with tumbling at room temperature and then was stopped by adding 1.5 mL of 1M glycine. Fixated cells were pelleted by centrifugation at 300g and frozen in liquid nitrogen prior to further processing.

1b - Sampling and fixation of mouse embryos:
For each 4C-seq sample, 10 whole E9.5 mouse embryos were incubated at 37°C for 45' with shaking in 500μL of a 0.125% collagenase solution (Roche, 10103578001) in PBS and then mecanically disrupted with a blue pipette tip. The resulting cell suspension was then filtered through a cell strainer with a mesh diameter of 65μm in order to discard tissue aggregates and fixed in 10 mL of a 2% solution of PFA in PBS. The fixation conditions and procedure were the same than the explained above for zebrafish.

1c - Sampling and fixation of amphioxus embryos:
Syncronized amphioxus embryos were sampled at either the 8hpf, 15hpf or 36hpf stages (8000 embryos per experiment in the case of 8hpf embryos, 4000 embryos in the other two cases). They were then fixed in 1.5 mL of a 1.85% PFA solution in MOPS buffer (0.1M MOPS pH 7.5, 2mM $MgSO_2$, 1mM EGTA and 0.5M $NaCl$). 155μL of 10% glycine was then added in order to stop the fixation, followed by several washes in NaPBS (PBS supplemented with $NaCl$ reaching a 0.5M concentration). Cell pellets were then frozen with liquid nitrogen.

1d - Sampling and fixation of centipede embryos:
250 germ-band *Strigamia* embryos were sampled. The chorions were pricked in order to make the embryo cells accessible and the yolk was solubilized in 1mL of Ginzburg Fish Ringer buffer. Deyolked embryos were collected by centrifuging at 300g for 30s and fixated in a 10mL solution of 2% PFA in PBS for 10' at room temperature. The fixation was stopped with 1.5 mL of 1M glycine and the cells were pelleted by centrifuging for 8' at 300g and frozen in liquid nitrogen for further processing.

2 - Isolation and permeabilization of nuclei:
Frozen pellets of fixated cells were resuspended in 5mL of ice cold Lysis Buffer (10mM Tris

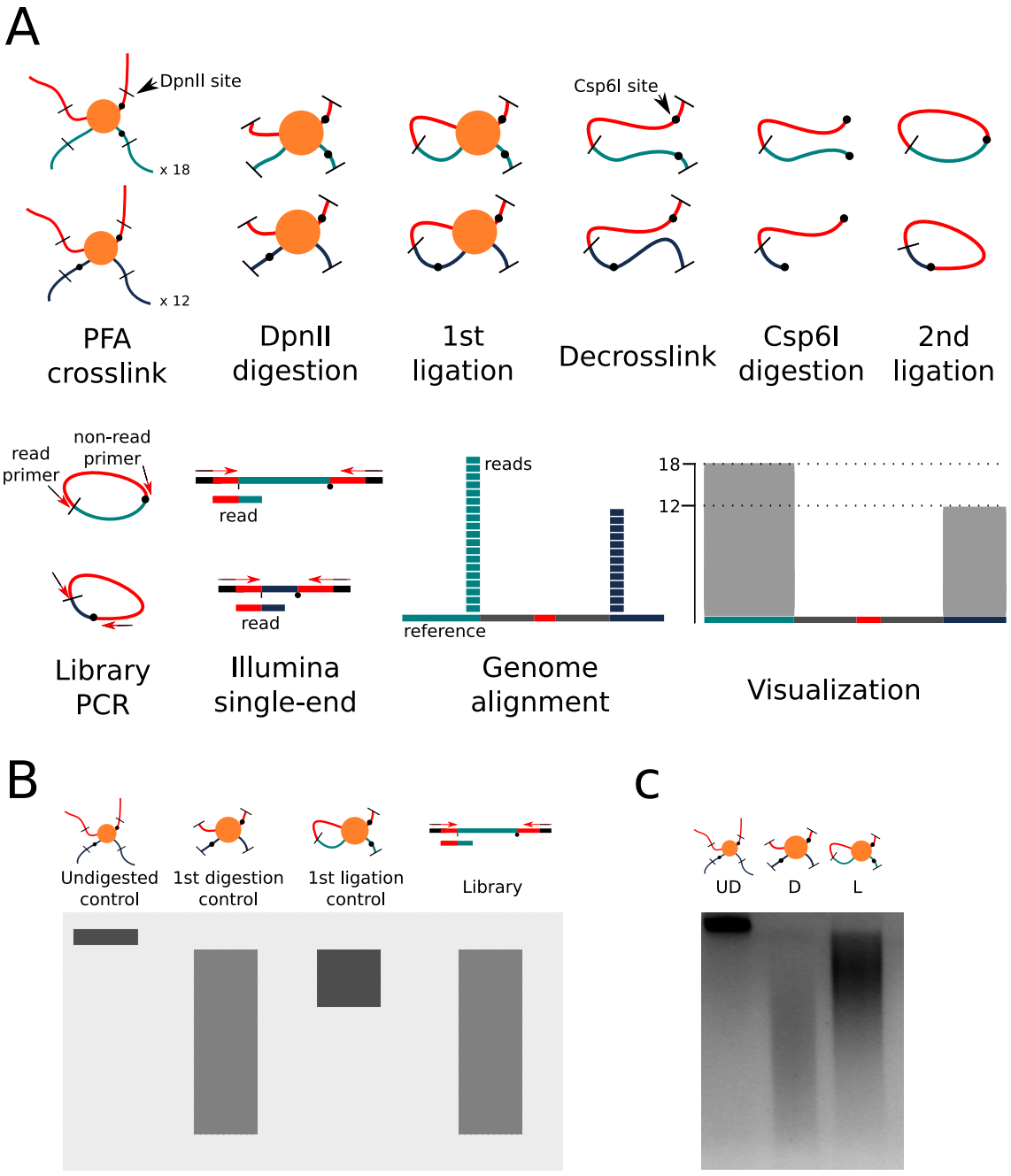


Figure 3.1: In (A) there is an graphic scheme of the 4C-seq protocol. The red fragment is the viewpoint and the turquoise and dark blue fragments are two different interacting loci. In this simplified example, the turqoise fragment is found together with the bait 18 times while the dark blue is found 12 times. The orange circles represent proteins mediating the interactions. After the first digestion and ligation, the red locus is joint together with either of the two other fragments. After the second digestion and ligation, circular ligation products are generated. Then, it is possible to amplify and sequence the interacting loci with specific primers designed in the sequence of the bait and containing the Illumina adaptors (in black). 50 bp single-end reads then contain both the primer corresponding to the bait fragment (again in red) plus around a 30 bp sequence from the interacting locus. The sequence of the primer is used to identify the different 4C-seq experiments and then is trimmed in order to align the remaining part to the reference genome. Then, the reads covering each loci are quantified and this reflects the proximity of each of them to the viewpoint. The final quantification is usually represented with barplots with the genome coordinates in the x axis and the number of interactions with the viewpoint in the y axis. In (B) there is and schematic representation of the expected agarose gels for the different quality controls performed. In (C) there is an actual example of one of the quality control gels: Undigested (UD), Digested (D) and Ligated (L).

pH 8, 10mM NaCl, 0.3% Igepal CA-630 (Sigma-Aldrich, I8896), 1x protease inhibitor cocktail (Complete, Roche, 11697498001)) and transferred to a 15mL tissue grinder (Tenbroeck Tissue Grinder, Wheaton 357426). The cells were lysed by moving the pestle of the homogeneizer up and down repeatedly with pauses to let the tissue cool down on ice. The state of the lysis was determined by taking several 3μL samples, staining them with the Methyl Green-Pyronin dye (Sigma-Aldrich, HT70116) and examining it in a microscope. The dye stains the cytoplasms in pink and the nuclei in green, and therefore if the lysis is ready many green nuclei and few pink cytoplasms should be observed. Once the lysis were ready, the nuclear suspensions were transferred to 15mL Falcon tubes and the isolated nuclei were pelleted by centrifugation at 600g. The nuclei were then resuspended in 500μL of 1x DpnII buffer (NEB, B0543). Then, 15μL of 10% SDS was added in order to permeabilize the nuclei and the sample was incubated at 37°C with shaking for one hour. The reaction was stopped by adding 75μL of Triton X-100 and the sample was again incubated for one hour at 37°C with shaking. Nuclear chromatin was then accessible for the subsequent steps of digestion and ligation.

3 - First digestion with primary restriction enzyme:

The next step was digesting the chromatin with a frequent cutter restriction enzyme, in our case DpnII. The reaction was set by adding to the permeabilized nuclei 10μL of 10x DpnII buffer plus 400U of the DpnII restriction enzyme (8μL from a 50U/μL batch, NEB R0543M). The reaction was incubated overnight at 37°C with shaking. Digestion efficiency was checked by taking two samples of 5μL before and after adding the DpnII enzyme. Those samples were decrosslinked by adding 90μL of Tris pH 7.5 10mM and 5μL of Proteinase K 10mg/mL (Roche, 03115844001) and incubating them for two hours at 65°C. They were then phenolized and loaded into a 0.8% agarose gel in order to check the size distribution of the DNA fragments. A single high molecular weight band is expected in the lane of the undigested sample, while a smear with DNA fragments of different sizes (ranging from several kilobases to 100-200 basepairs) is expected after DpnII digestion (Figure 3.1B and Figure 3.1C).

4 - First ligation (capture of interacting fragments):

The digestion reaction was stopped by heat inactivation for 20' at 65°C. Then the ligation reaction was set in a 50mL Falcon tube (5.7mL Milli-Q water, 700μL of 10x ligase buffer (Thermo scientific, B69), 12μL of 5U/μL T4 DNA ligase (Thermo scientific, EL0014)). With this ligation, restriction fragments that could be potentially far in the linear DNA sequence but close in 3D are linked together consecutively. The ligation reaction is set in a big volume so that random ligations are disfavored because of the dilution. After the capture, the objective is to identify and quantify the junctions between restriction fragments. The ligation was kept overnight at 16°C. Ligation efficiency was also checked taking a 100μL control, decrosslinking it and loading it into a 0.8% agarose gel together with the undigested and digested controls obtained in the previous step. A shift towards higher molecular weight DNA fragments is expected in the lane with the ligation in comparison with the lane with the digestion (Figure 3.1B and Figure 3.1C).

5 - Descrosslinking and purification of ligation products:

Now that the 3D interactions are captured proteins are no longer needed and it is the time to purify the ligation products. For that purpose, first the sample is decrosslinked by adding 30μL of 10mg/mL Proteinase K and incubating it overnight at 65°C. Then, 30μL of RNAse

10mg/mL was added and the sample incubated at 37°C for 45' in order to eliminate the RNA. After that the sample was phenolized and the DNA precipitated with ethanol at -80°C for three hours (35mL of pure ethanol, 7mL of Milli-Q water, 1mL of 3M NaAc pH 5.2, 7μL of Glycogen 1mg/mL). The DNA was pelleted by centrifuging the ethanol mixture for 1 hour at 3900g, then washed once with 10mL of 70% ethanol and eventually resuspended in 445μL of Milli-Q water.

6 - Second digestion with secondary restriction enzyme:

To identify juntions involved with particular loci of interest, circular ligation products are needed in order to PCR enrich ligation events including the restriction fragment of interest. For that reason, purified ligation products were digested again with a second restriction enzyme, Csp6I, and recircularized with a second ligation step. 50μL of 10x Thermo Scientific Buffer B (BB5) and 5μL of 5U/μL Csp6I enzyme (Thermo Scientific, ER0211) were added to the 445μL sample from the previous step. Digestion was left overnight at 37°C.

7 - Second ligation (circularization of ligated fragments):

The digestion reaction was stopped again by heat inactivation for 20' at 65°C. Then, the second ligation reaction was set, again in diluted conditions (12.1 mL of Milli-Q water, 1.4mL of 10x ligation buffer and 20μL of 5U/μL T4 DNA ligase). The ligation was incubated overnight at 16°C.

8 - Purification of ligated circular products:

Circularized ligation products are then purified by dyalisis using Amicon Ultra-15 10KNMWL centrifugal filters (Millipore, UFC901024). Briefly, the whole ligation is introduced in the centrifugal filter and centrifuged at 4000g for 15'. The flowthrough is discarded and the sample is washed twice with 15mL of Tris pH 7.5 10mM. Finally, when the volume that remains in the upper part of the centrifugal tube (that contain the concentrated circularized ligation products) is approximately 500μL, the sample is recovered, transferred to a clean tube and quantified using the Qubit dsDNA HS Assay (Thermo Scientific, Q32851).

9 - 4C-seq primer design and library preparation:

Finally, different single-end Illumina libraries were prepared for the different viewpoints or baits of interest using PCRs with specific primers also containing the Illumina single-end tails. The sequence and location of the primers of the different 4C-seq experiments are available in a tsv table in the gitlab repository (acemelthesis/4C-seq/primers_list.tsv). It is important to note that the experiments for the zebrafish *hoxd10a*, *hoxd11a* and *hoxd13a* promoters were taken from a previous publication (Woltering et al. 2014) and therefore the information of the primers is not included. The design of the primers is similar to the design of an inverse PCR. First, the restriction fragments that were going to be used as baits were chosen using the following criteria: they should be DpnII/Csp6I fragments close enough to the loci of interest and at least 300bp long. Then, the primer carrying the Illumina read adaptor is designed strictly overlapping the DpnII restriction site and pointing outwards from it, ensuring that most of the sequencing read will contain not the sequence of the bait itself but the sequence of the different unknown ligated fragments. The non-read primers design is more flexible, but these primers also need to point outwards from the bait to the opposite direction with respect to the read primer. For each library, eight individual and equivalent 25μL PCR reactions were pooled in order to minimize PCR biases. The polymerase chosen was the

Expand Long Template PCR System (Roche, 11759060001), because the range of amplicon sizes in unknown *a priori* and the influence of the size of the restriction fragments ligated to our baits must be minimized. The PCR mixes were then prepared as follows:

- 2.5 μL of 10x PCR buffer (Expand Long Template)
- 0.5 μL of 10x dNTPs
- 2 μL of 10mM primers mix
- 0.35 μL of Expand Long Template Polymerase
- x μL of template (as needed for at least 50ng total)
- x μL of Milli-Q water up to 25μL of total reaction

The PCR program used was the following, based on the recommendations for the specific polymerase:

Initial denaturalization	94°C	2'	
Denaturalization	94°C	10"	Up to 35 cycles
Annealing	50-65°C	1'	
Extension	68°C	3'	
Final extension	68°C	10'	

Several parameters such as the annealing temperature, the number of cycles or the amount of template required were optimized for each particular experiment in single reactions that were then loaded in 1.5% agarose gels. Finally, the 8 final PCRs for each of the experiments are pooled and purified using the AMPure XP beads protocol (Beckman Coulter, A63882).

3.1.2 4C-seq data analysis

We will divide the explanation of the analysis in two. In the first part, we will describe how 4C-seq data is processed from raw data to data visualization. In the second we will focus on the downstream analysis: how significant interactions are called, how RLs are called and how the calculations for the different 4C-seq related figures were done. The code used will be included in the gitlab repository and referenced in this section.

(a) From raw fastq files to contact visualization:

The processing of the data is automatized with two scripts, a Python2 script that classify the reads from the sequencing lane in different files for the different baits and a Perl script that pick those reads and process them until they can be visualized in a genome browser. Those scripts and example files to test them are available in the gitlab repository. In addition, instructions about how to run the example analysis are also included in the following file: acemelthesis/4C-seq/1-Raw2Visual/raw2visual.md.

1. *Demultiplexing*:
   Sequencing lane files contain 50bp single cell reads from several 4C-seq experiments with different baits. The beginning of the read correspond to the sequence of the primer in the bait followed by the sequence of the interacting fragment. Using the sequence of

the primer it is possible to identify the reads coming from the different experiments and split the reads in different fastq files. The sequence of the primer is also removed so that only the parts of the reads coming from the different interacting loci are kept and are then alignable to the reference genome.

2. *Alignment*:
   Bowtie1 mapper was used to align the different experiments to their respectives reference genomes. For zebrafish, the *danRer7* assembly was used in the Hox section of the results (4.1, p.75) while *danRer10* was used for the global changes in architecture section (4.2, p.88). The rest of the assemblies used were *mm9* (mouse), *Smar1* (*Strigamia maritima*) and the recently published *Bl71nemr* for the european amphioxus (Marlétaz et al. 2018). Reads mapping to more than one location were filtered out.

3. *Filtering reads*:
   From the fasta of the reference genome two restriction maps are generated in bed format. One of them contain all the DpnII-DpnII restriction fragments and the other contain all the valid DpnII-Csp6I fragments. Valid restriction fragments are those that are flanked by one of each of the restriction sites and are bigger than 40bps. Those fragments that are closer than 5kb to the viewpoint are also filtered out. Then, the aligned reads that overlap with good restriction fragments are maintained and the rest are discarded (Figure 3.2A).

4. *Scoring of restriction fragments*:
   Then, the number of filtered reads per DpnII-DpnII restriction fragments are quantified. This number is divided by two if the two fragment ends are valid, and left unchanged if only one of them is (Figure 3.2A). Before visualization, the data is smoothened using a sliding window of 31 fragments. Then, the score of each of the fragments is converted to the mean score of the fragment plus the 15 fragments surrounding it (see Figure 3.2B, the top two panels show the raw and the smoothed 4C-seq signal).

(b) Downstream analysis:

1. *Interaction calling (Poisson background)*:
   For the results of the section 4.1 (page 75), significant interactions were identified using Poisson distributions as background from the smoothed files. The read proportions given in that section are always from reads located inside significant interactions. Briefly, a different lambda (λ) parameter is inferred for each 4C-seq experiment. In order to do that, the scores of each of the restriction fragments are shuffled 10 times, and the scores of these randomized experiments are then smoothened using the sliding window. Then, the scores corresponding to the 95th percentile of each of the experiments are saved and the maximum of them is used as the final value of λ (third panel of Figure 3.2B). Restriction fragments with scores significantly above the λ background are merged and considered interactions (fourth panel of Figure 3.2B). This interaction calling is performed automatically by the same Perl script that wraps up the processing from the fastq files to the bedgraphs ready to visualize (see instructions at acemelthesis/4C-seq/1-Raw2Visual/raw2visual.md). In addition to that, the code in R used to illustrate the interaction calling in Figure 3.2B is also available (check acemelthesis/4C-seq/2-DownstreamAnalysis/interaction_caller/intcaller_sample.rmd).

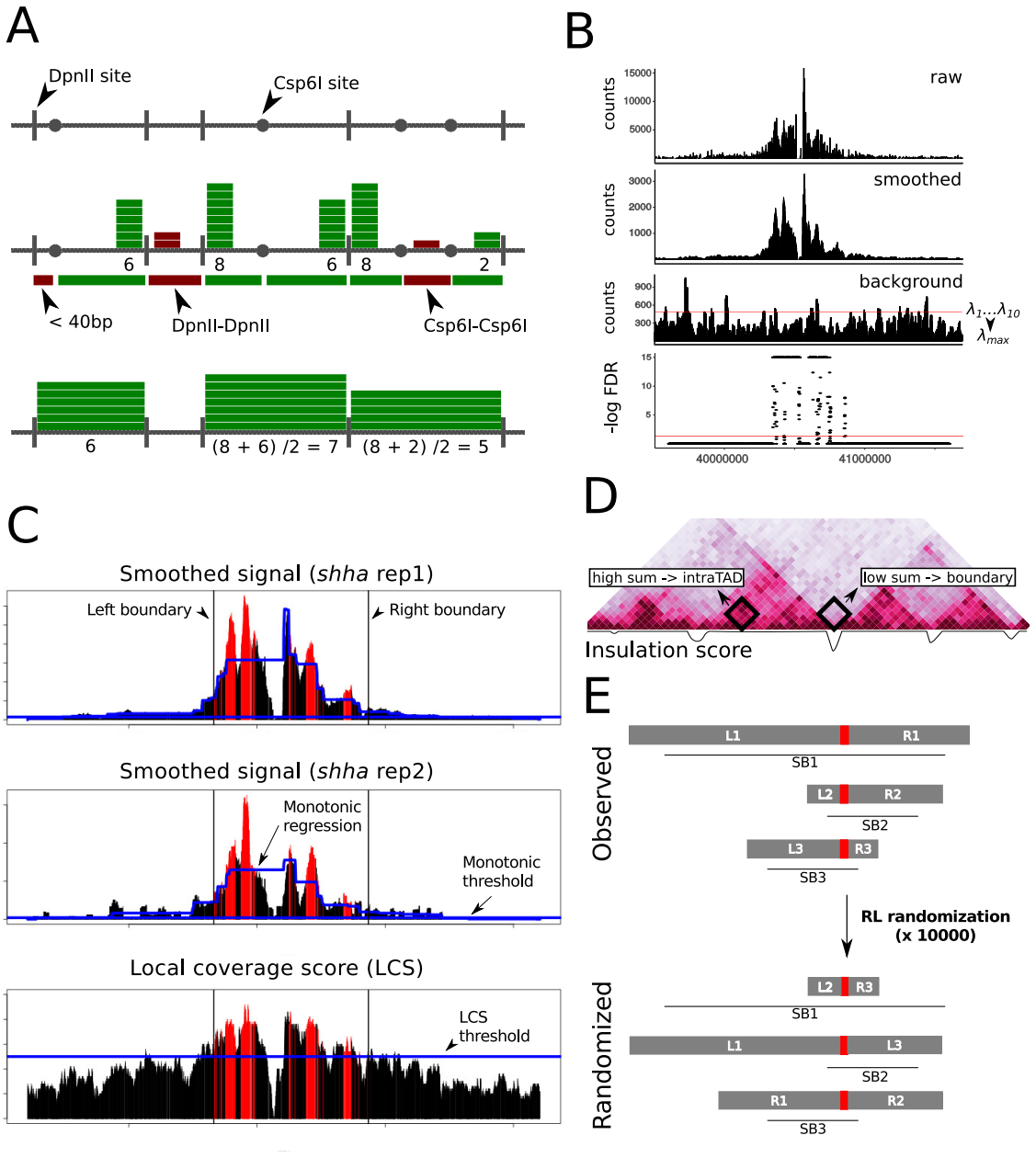


Figure 3.2: The process of read filtering and restriction fragment scoring is illustrated in (A). Good DpnII-Csp6I fragments are marked with green boxes while non informative ones are marked with dark red boxes. Reads mapping in "bad" fragments are also painted red and are discarded. The four DpnII-DpnII fragments are scored as follows: in the first one, only the right end of the fragment is informative and 6 reads map on it, then the score is 6; in the second there is no Csp6I restriction site in the middle so the count is 0; in the last two fragments both ends are informative and then the mapped reads are normalized dividing by 2. In (B) the smoothening and the interaction calling of the 4C-seq profiles are illustrated. In the top panel, the raw interaction profile of *shha* is shown as an example. In the second one, the smoothed profile of the experiment above is shown, obtained using a 31 fragments sliding window. In the third panel, an individual randomization of the profile used for the calculation of the Poisson λ is depicted. The red line represents the 95^{th} percentile, that corresponds to the value of λ in the case that this value is the highest of all the 10 randomizations. Finally, in the bottom panel the FDR values obtained from the Poisson test for the different restriction fragments are shown. The red line represents the FDR=0.05 cutoff. In (C) the landscape caller is explained. The top two panels are the two replicates of a 4C-seq experiment and the bottom one the Local Coverage Score. In red the significant interactions are highlighted, that are those that are consistently above the monotonic regression (blue lines in the top two panels) and with an LCS value above the threshold (blue line in the bottom panel). In (D) there is an schematic of the calculation of the insulation score in HiC matrices. The insulation score in each point is the sum of the depicted sliding squares, being the sum high in intraTAD loci and low in boundary elements. Finally, in (E) is shown how the RLs are randomized with respect to the sytenic blocks (SB) in order to calculate empiric p-values for the overlap between both features.

2. *Regulatory Landscape calling*:
   For the results of the first part of the section 4.2 an automatic prediction of the extension of the RLs from the 4C-seq experiments was carried out. This was performed by calling interactions differently, using a monotonic regression as a background and rank products in order to integrate the information from several replicates as proposed by Geeven et al. 2018 with slight modifications (as performed in Marlétaz et al. 2018). The upstream most and the downstream most significant interactions are considered the boundaries of the RL. Briefly, the raw 4C-seq profiles were smoothed using a 51 fragment window. Then, using the peakC R package (Geeven et al. 2018), two monotonic regressions are fitted at each side of the viewpoint for both replicates. The actual interaction scores of each of the restriction fragments are then compared with the monotonic background and initial significant interactions are identified using the rank product statistics. Then, these interactions are further filtered according to different criteria. First, interactions happening in poorly covered regions are eliminated. In order to do that, a local coverage score (LCS) is calculated for each fragment that consist in the number of fragments surrounding it covered by at least one read (the default window is also 51 fragments). The LCS of a fragment must be in the top 25% of the experiment to be eligible as an interacting fragment. In addition, the value of the monotonic regression in interacting fragments must be sufficiently high to ensure an adequate signal to noise ratio. In this work, the monotonic regression value in interacting fragments must be above the 2.5% of the value of the monotonic regression at the viewpoint. This processing is illustrated in Figure 3.2C. The process is automated by an Rscript available in the gitlab repository. Detailed information on how to replicate the analysis can be found as well there (acemelthesis/4C-seq/2-DownstreamAnalysis/landscape_caller/landscape_call.md).

3. *Regulatory Landscape validation using HiC*:
   In order to validate the RL boundaries calculated we went to compare the predictions performed in mouse with the publicly available HiC data from mESC (Dixon et al. 2012). In order to do that we plotted the overlay of the HiC signal 1Mb around all the predicted RL boundaries at 40kb resolution. We repeated this plot with the signal around the boundaries of equally sized randomized genomic chunks, incorporating the signal from 100 independent randomizations. In addition, we also calculated an insulation score equivalent to the one calculated by Crane et al. 2015 using a 200kb window of both the real and the randomized overlays. An schematic of how this score is calculated is shown in Figure 3.2D. The resulting plots are shown in the corresponding results section (4.2.1, in the Figure 4.7B of p.92). The Python3 code used to perform this analysis is available in the gitlab repository (acemelthesis/4C-seq/2-DownstreamAnalysis/landscape_validation/RLvsHiC.ipynb).

4. *Regulatory Landscape size and enhancer content comparison*:
   The comparison of RL sizes between different species is trivial using the previously calculated and validated boundaries. ATAC-seq peaks from stages equivalent to the 4C-seq experiments were used as a proxy to enhancer content, and are those from Marlétaz et al. 2018. They were intersected with the RLs by using bedtools intersect with the -c flag. The R code to produce the plots in Figure 4.7C is available in the gitlab repository (acemelthesis/4C-seq/2-DownstreamAnalysis/landscape_size/fourC_lsize.Rmd).

5. *Regulatory Landscapes relationship with syntenic blocks*:
   In the section 4.2.2, the calculated RLs from zebrafish are compared with the syntenic blocks involving the bait sequences (Figure 4.8B, p.94). The calculation of those syntenic blocks are explored later in this chapter (subsection 3.3.1, p.72). Here we will cover how the agreement between those syntenic blocks and RLs is quantified. First, the overlap between the syntenic blocks and RLs is calculated in both directions (proportion of the RL covered by the syntenic block of the bait and viceversa). Then, the extension of the RLs is randomized 10,000 times and the overlaps are calculated again in order to determine the expected overlap by random chance and get empiric p-values. Instead of choosing random genomic positions, since our RLs are enriched for developmental genes, we chose to randomize the extension of the RLs but centered on the same loci. In order to do that, we took the upstream and downstream extension of all the RLs and placed them randomly around the existing baits (see scheme in Figure 3.2E). The Python3 code used to generate those randomizations and the subsequent plots is available in the gitlab repository (acemelthesis/4C-seq/2-DownstreamAnalysis/landscape_synteny/RLvsSynteny.ipynb).

3.1.3

3D modelling of chromatin from 4C-seq experiments

In order to produce the 3D-models of the Hox clusters of zebrafish, amphioxus and *Strigamia maritima* the *4Cin* pipeline was used (thoroughly described in Irastorza-Azcarate et al. 2018, including a link to the github repository). These are the results presented in the sections 4.1.2 and 4.1.4. Here we will just describe the rationale and present some analysis that highlights the reliability of the method.

4Cin models the chromatin as a string of beads of different sizes representing an equal count of restriction fragments (see Figure 3.3A). The position of these beads is then optimized according to the restraints imposed by the 4C-seq experiments, assuming that higher interaction frequencies correspond with smaller distances in the 3D space. This optimization is done using the IMP software (Russel et al. 2012) and thousands of models are generated and then clustered. Eventually, a consensus model is finally extracted and the distances can be represented as a heatmap that resemble those of HiC experiments and were named v-HiCs.

Several controls were made in order to validate the v-HiC heatmaps. *4C-in* was tested using available 4C-seq experiments performed in mice and the resulting v-HiC maps were compared with available HiC experiments from mouse ES-cells (Dixon et al. 2012). Highly similar contact matrices were obtained in the three different loci explored (Figure 3.3C). In addition, the method was robust to the supression of 4C-seq baits. In the case of the amphioxus Hox locus, up to 8 out of the total 14 could be removed with minor differences in the final output (the correlation dropped less than 0.2 points). In the case of the zebrafish hoxD locus, 6 out of the 9 could also be retrieved also with minor deviations from the model using the whole set of baits (3.3B). Again, further discussion on the *4Cin* algorithm and the validation can be found in Irastorza-Azcarate et al. 2018.

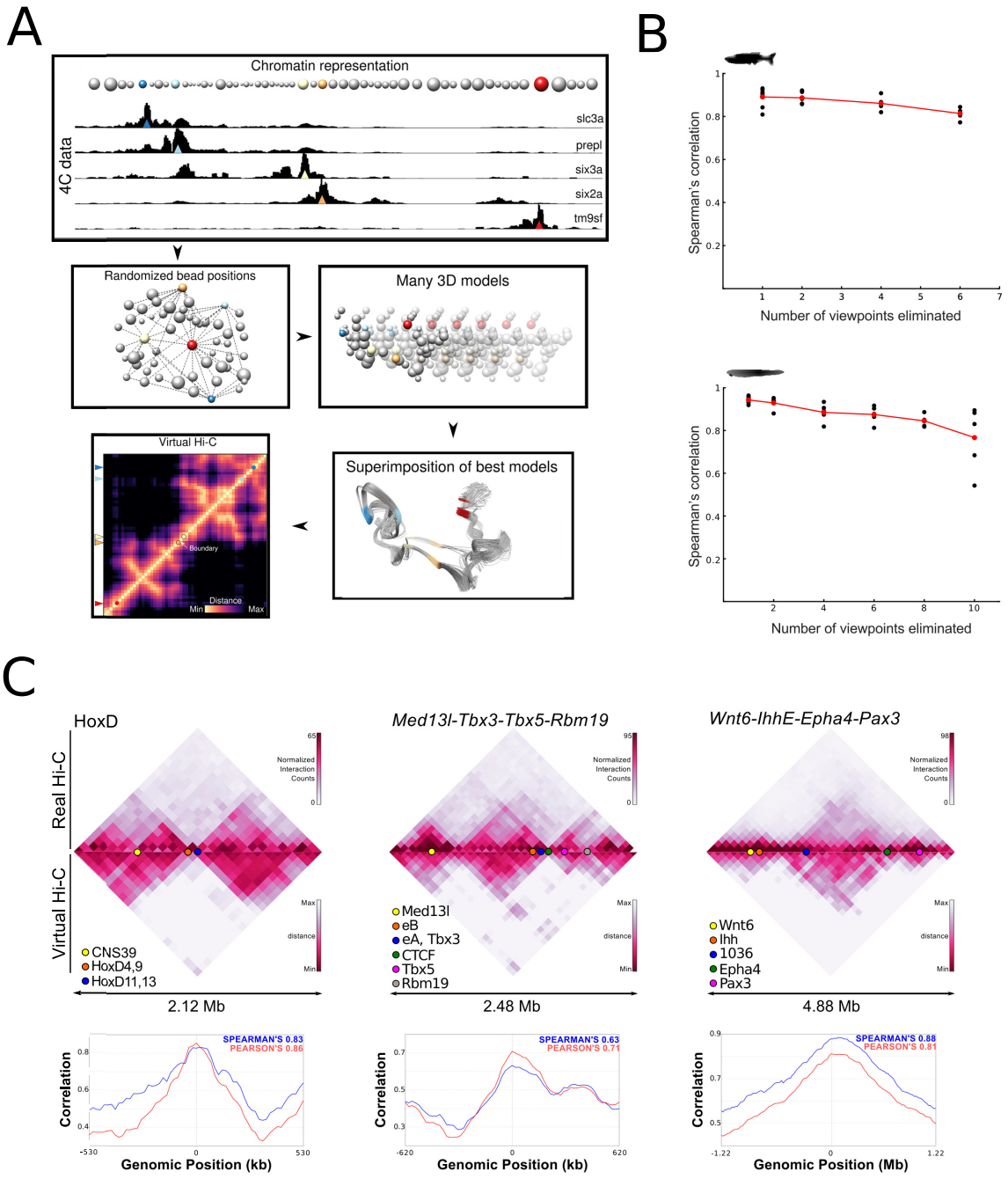


Figure 3.3: Modelling of chromatin using 4C-seq restraints. (A) Workflow of the 4Cin modelling method. (B) Robustness of the model to the reduction of the number of 4C-seq viewpoints used as restraints. The top plot shows the correlation of the subsampled models in zebrafish with respect to the original model with the whole set of viewpoints. The bottom plot is the equivalent for amphioxus. (C) Comparison of 4Cin generated models and v-HiC matrices in mouse with the corresponding HiC experiment performed in mouse ES cells (Dixon et al. 2012). The correlation is calculated with the bins matching and also sliding the bins. The maximum correlation is always reached when the bins between the model and the real HiC experiment matches.

3.2

HiChIP and HiC

Similarly to the case of 4C-seq, the rationale of the techniques are explored in the Introduction (1.3.1, p.15). Here we will describe the HiChIP protocol for several histone modifications and the

data analysis procedures for both this technique and HiC.

3.2.1

HiChIP library preparation

Being another flavor of the C-technique, the HiChIP protocol is partially overlapping with the 4C-seq one and mostly follows the one described in Mumbach et al. 2016. An schematic cartoon of the protocol is shown in Figure3.4A. In this work, we performed HiChIP experiments using the H3K4me3, H3K27ac and H3K27me3. We did so both in zebrafish embryos of 24hpf and 80% epiboly and in amphioxus embryos of 15hpf.

1a - Sampling and fixation of zebrafish embryos :
The sampling procedure is equivalent to the explained before for the 4C-seq (see 3.1.1, p.51) and the only difference in the fixation was that embryos were fixed in 1% PFA PBS solution. In this occasion, 1,000 whole embryos were used for the 24hpf experiments and 3,000 for the 80% epiboly.

1b - Sampling and fixation of amphioxus embryos :
Fixation of amphioxus embryos was equivalent to the 4C-seq fixation (see 3.1.1, p.51).

2 - Isolation and permeabilization of nuclei :
Cell lysis was performed as in the 4C-seq experiments. However, nuclear permeabilization differed. After the 4°C centrifugation to pellet the nuclei they are resuspended in 100μL of 0.5% SDS and incubated at 62°C for 10'. Then the SDS was quenched by adding 292μL of Milli-Q water and 50μL of Triton X-100 10%. 5μL of the sample are reserved at this point in order to check chromatin integrity.

3 - Chromatin digestion with 4 cutter restriction enzyme :
The chromatin was then readily digested using the DpnII restriction enzyme (NEB, B0543). 50μL of DpnII buffer 10x and 8μL of 50U/μL of enzyme were used (200U total). The mix was incubated for 2 hours at 37°C with rotation. After that, 5μL of the digestion were taken as control. This sample was decrosslinked and run in a 1.5% agarose gel together with the sample taken before adding DpnII. The interpretation of this gel is equivalent to the interpretation of the 4C-seq gels (see Figure3.4B and 3.4C).

4 - Fill in of cohesive ends including biotinylated ATP :
After the gel check, the reaction was stopped by heat inactivation at 65°C for 20'. 10μL of the digestion were reserved at this point. Then, the cohesive ends left by the DpnII digestion were filled using Klenow (NEB, M0210) incorporating a biotinylated version of adenine. This was done by adding 50μL of the Incorporation Master Mix: 36μL of biotin-dATP 0.4mM (Thermo Fisher, 19524016), 4μL of dNTP mix 10 mM (excluding dATP) and 10μL of the Klenow enzyme (5U/μL). The biotinylated nucleotide will allow to enrich the library in ligation junctions later. The incorporation reaction is incubated for 45' at 37°C with rotation.

5 - Blunt ligation (capture of interacting fragments) :
After the fill-in, the ligation reaction is set. Equivalently to the 4C-seq experiment, the

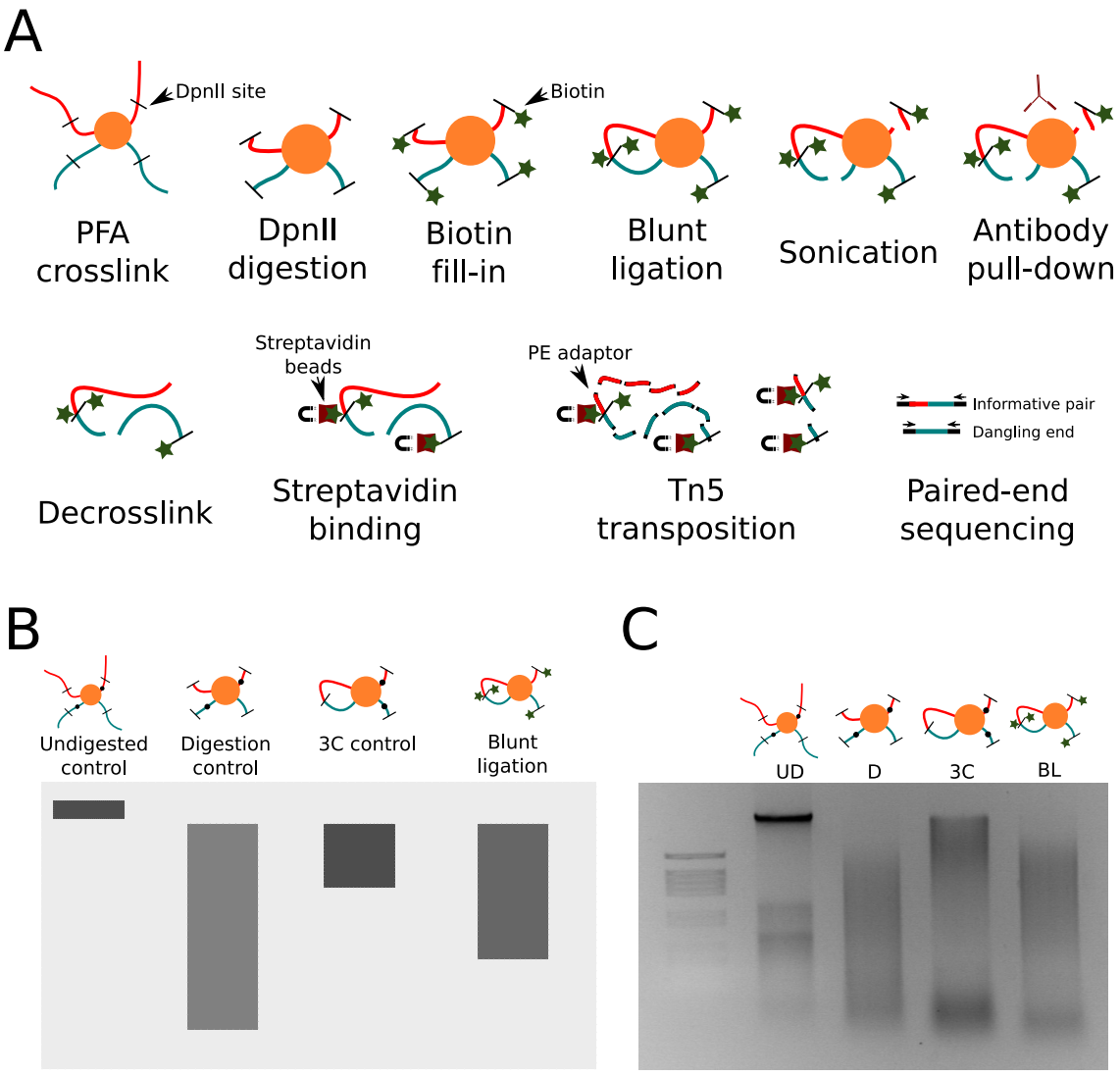


Figure 3.4: In (A) there is a graphic scheme of the HiChIP protocol. In the example, the red fragment is interacting with the turquoise one in the vincinity of a given protein (orange circle). The interaction is stabilized with PFA fixation, the chromatin is digested with DpnII and the ends are filled with biotin modified adenine nucleotides. After the blunt ligation, the red and the turquoise fragment become consecutive in the linear sequence. Then there is a sonication step and an antibody pull-down that select those interactions that happen near an specific protein (in the case of this work we use histone modifications). After the decrosslink, the ligation events are enriched by using magnetic streptavidin beads. Then adaptors are introduced by Tn5 transposition and the beads are washed so mostly ligation events are retained for paired end sequencing. In (B) there is and schematic representation of the expected agarose gels for the different quality controls performed. In (C) there is an actual example of one of the quality control gels: Undigested (UD), Digested (D), 3C control (3C) and Blunt Ligation (BL).

ligation allows to capture in the same molecule restriction fragments that are close in the 3D space. In this case, however, the ends of the restriction fragments are blunt and therefore the ligation less efficient. Briefly, 950μL of the ligation master mix is added to the sample and incubated overnight at 16ºC with shaking. The Ligation Master Mix is composed by 670μL of Milli-Q water, 150μL of 10x NEB T4 DNA ligase buffer (including 10mM ATP, NEB BO202),125μL of Triton X-100 10%, 15μL of BSA 10mg/mL and 10μL of T4 DNA ligase (400U/μL, NEB M0202). An equivalent 90μL mix was prepared to perform a cohesive

ligation with the 10μL digestion sample that was not used in the fill-in reaction (termed 3C control). After the ligation overnight those samples were decrosslinked and run in a gel to check the ligation efficiency. This should be high in the case of the 3C control (with an obvious shift towards big DNA fragments when compared to the digestion control) and milder in the actual blunt ligation (Figure3.4B and 3.4C).

6 - Nuclear lysis and chromatin sonication :

After checking the ligation efficiency, the nuclei were pelleted again by centrifugation (5' at 2500g). The supernatant was removed and nuclei were resuspended in 1mL of Nuclear Lysis Buffer: Tris-HCl pH 7.5 50mM, EDTA 10mM, SDS 1%, 1x protease inhibitor cocktail (Complete, Roche, 11697498001). They are kept for 5' on ice before adding 2mLs of ChIP Dilution Buffer: Tris-HCl pH 7.5 16.7mM, EDTA 1.2mM, NaCl 167mM, SDS 0.01%, Triton X-100 1.1%. Then the chromatin was split in three 1mL aliquots and sheared using the Covaris M220 sonicator with the following parameters: duty cycle 10%, PIP 75W, 100 cycles/burst, time 5'. Sonicated chromatin was then centrifuged for 15' at 16,000g and the supernatant was transferred to a new tube. Before proceeding to chromatin immunoprecipitation, sonication efficiency was checked taking a 20μL aliquot. This aliquot was treated with 1 μL of RNAseA 10mg for 30' at 37°C, then decrosslinked with 1μL of proteinase K 10mg/mL for 1h at 65°C, phenolyzed and loaded in a 2% agarose gel. The expected median size of the sonicated DNA fragments should be around 300bps.

7 - Chromatin immunoprecipitation :

In this step, using a specific antibody, we will select those interactions that happen in the vincinity of nucleosomes with specific histone tail modifications. Before adding the corresponding antibodies, samples are precleared treating them with Dynabeads Protein G magnetic beads (Invitrogen, 10003D). This treatment ensures that chromatin that bind to the beads inespecifically (without antibodies involved) is washed out. Briefly, 90μL of the beads were washed twice with 1mL of ChIP Dilution Buffer (check step 6) using the magnetic rack. Then they were resuspended in 150μL of ChIP Dilution Buffer and 50μL was added to each of the three tubes with the samples and the mix was incubated for 1h with rotation at 4°C. After the incubation the beads were reclaimed and the supernatants with the precleared samples transferred to a new tube. Then, a total of 20μg of antibody per experiment was added, either anti-H3K4me3 (abcam, ab8580), anti-H3K27ac (abcam, ab4729) or anti-H3K27me3 (Millipore, 07-449). The antibodies were incubated with the chromatin overnight at 4°C. Then, the chromatin bound by the antibodies was captured with the Dynabeads Protein G. Again, 90μL of the beads were washed twice with 1mL ChIP Dilution Buffer and resuspended in 150μL of the same buffer. Then, 50μL of the beads were added to each sample tube and incubated for 2h at 4°C to pull down the antibody. Then, the beads were washed three times with 1mL of three different buffers (nine washes total) with increasing salt concentration. The composition of the three different buffers was the following: Wash Buffer 1 (Tris-HCl pH 7.5 20mM, EDTA 2mM, NaCl 150mM, SDS 1%, Triton X-100 1%), Wash Buffer 2 (Tris-HCl pH 7.5 20mM, EDTA 2mM, $NaCl$ 500mM, SDS 0.1%, Triton X-100 1%) and Wash Buffer 3 (Tris-HCl pH 7.5 10mM, EDTA 1mM, $LiCl$ 250mM, Igepal CA-630 (Sigma-Aldrich, I8896) 1%, Na-deoxycholate 1%). Then the sample is washed three times more with TE before the elution and reverse crosslinking.

8 - Chromatin elution and reverse crosslinking :

Pulled down chromatin was then separated from the magnetic beads by resuspending them in 150μL of Elution Buffer ($NaHCO_3$ pH 8.8 50mM, SDS 1%) and incubating the mix for 3' at 37°C with shaking. The supernatant is kept and the process is repeated with another 150μL of Elution Buffer. The merge of the supernatants contained the immunoprecipitated chromatin that was then ready for reverse crosslinking. For that purpose, 15μL of proteinase K 10mg/mL was added to each 300μL sample. They were incubated at 55°C for 45' and at 65°C for 1h 30' and then the DNA was purified using two DNA Clean and Concentrator columns (Zymo Research, D4013). Purified DNA containing ligation junctions was eluted separatedly in 10μL of Milli-Q water and quantified using Qubit dsDNA HS Assay (Thermo Scientific, Q32851).

9 - Biotin assisted pull-down of ligation junctions :

Then ligation junctions were enriched using streptavidin magnetic beads, taking advantage of the biotinylated adenine that was incorporated in the fill-in step. 5μL of Dynabeads My One Streptavidin C1 (Thermo Fisher, #65001) were washed with 500μL of Tween Wash Buffer (Tris-HCl pH 7.5 5mM, EDTA 0.5mM, NaCl 1M, Tween-20 0.05%) and resuspended in 20μL of Biotin Binding Buffer (Tris-HCl pH 7.5 5mM, EDTA 0.5mM, NaCl 1M). 10μL of the beads were added to each of the HiChIP samples and incubated for 15' at room temperature with shaking. The junctions including the modified nucleotides were attached then to the magnetic beads. Then the beads were washed twice with 500μL of Tween Wash Buffer, including an incubation step with the buffer of 2' at 65°C. Then, the Tn5 transposase (CABD Proteomic Service, produced as proposed in Picelli et al. 2014) is used to add Illumina adaptors. For that purpose, the beads were first washed with 100μL of TD buffer 1x (Tris-HCl pH 7.5 10mM, $MgCl_2$ 5mM, Dimethylformamide 10%). Then they were resuspended in 25μL of 2x TD buffer. Tn5 and Milli-Q water up to a total volume of 50μL is then added. The amount of Tn5 to use depends on the amount of DNA quantified at the end of the step 8, and ranges from 0.1μL to 4μL. If the amount of DNA recovered is quantifiable by Qubit, a proportion of 2.5μL of Tn5 per 50ng is added. If recovered DNA is too low, then 0.1μL of enzyme is used. The transposition reaction is incubated for 10' at 55°C with shaking. Then, the reaction is stopped by reclaiming the beads and substituting the Tn5 mix with 100μL of 50mM EDTA, incubating this for 30' at 50°C. Then, the sample is washed twice with 100μL of 50mM EDTA, then once with Tween Wash Buffer and twice again with Tris-HCl pH 7.5 10mM. Now the enriched ligated junctions are ready for the final library amplification.

10 - Library amplification and purification :

Magnetic beads can be used directly as template for the amplification of the library using Illumina primers. But first, the appropriate number of cycles needs to be inferred. In order to do that, first a 5 cycles PCR is performed using the following mix:

- 25μL of NEBNext High-Fidelity 2X PCR Master Mix (NEB, E6040L)
- 0.5μL of Nextera Ad1_noMX primer 25μM
- 0.5μL of Nextera Ad2.X primer 25μM
- 24μL of Milli-Q water

The program was then the following, according to NEBNext specifications:

Hot start	72°C	5’	
Initial denaturalization	98°C	30”	
Denaturalization	98°C	10”	5 cycles
Annealing	63°C	30”	
Extension	72°C	30”	

Then the beads were reclaimed and 2μL of the reaction were used as template for a RT-PCR to estimate the total number of cycles. The rest of the reaction was preserved at 4°C. The mix and the program were the following:

- 2μL of the previous reaction
- 4.5μL of NEBNext High-Fidelity 2X PCR Master Mix
- 0.4μL of Nextera Ad1_noMX primer 12.5μM
- 0.5μL of Nextera Ad2.X primer 12.5μM
- 1μL of SYBR 10x (Thermo Fisher, S7563)
- 1.7μL of Milli-Q water

Initial denaturalization	98°C	30”	
Denaturalization	98°C	10”	25 cycles
Annealing	63°C	30”	
Extension	72°C	30”	
Final extension	72°C	5’	

After estimating the number of cycles, the remaining reaction was put back in the thermocycler to finish the library amplification with a equivalent program:

Initial denaturalization	98°C	30”	
Denaturalization	98°C	10”	estimated cycles
Annealing	63°C	30”	
Extension	72°C	30”	
Final extension	72°C	5’	

Finally, the PCR product is purified using the DNA Clean and Concentrator columns (Zymo Research, D4013), eluted in 20μL of Tris-HCl pH 8 10mM and quantified using Qubit dsDNA HS Assay (Thermo Scientific, Q32851). The samples were then sent for Illumina Paired-End sequencing.

3.2.2

HiChIP data analysis

Again we will divide the data analysis explanation in two parts, one referring to the processing from fastq files to contact matrices and contact visualization and a second one dedicated to the downstream analysis of these contact matrices. The first part is equivalent for all kind of HiChIP experiments regardless of the antibody and is based in the *TadBit* pipeline (Serra et al. 2017). The second part includes among other things the automatic prediction of RLs based on H3K4me3 HiChIP matrices, the prediction of enhancer-promoter hubs from H3K27ac matrices or the analysis of RL sizes related to WGDs.

(a) From raw fastq files to contact visualization :

This part of the analysis is automatized by several Python scripts. The guide on how to use them is available in the gitlab repository (acemelthesis/HiChIP/1-Raw2Visual/raw2visual.md).

1. *Alignment*:
   Each Paired-End read file is mapped independently using GEM mapper (Marco-Sola et al. 2012), given that we need to allow read pairs containing sequences located far away in the linear distance. The parameters used for the mapping are the ones specified in TadBit that are the GEM defaults with the following exceptions: only one alignment is reported (−−max−decoded−strata 1; −−min−decoded−strata 0) and the edit operations allowed are 4% of the length of the read (besides mismatches, therefore allowing indels, −e 0.04). Aligned reads are associated to a given restriction fragment in the genome and paired *a posteriori* according to the read names.
2. *Filtering of valid read pairs*:
   Once the pairs of reads are associated to a given restriction fragment the filtering step can proceed by classifying the reads in categories (Figure 3.5A). Bear in mind that some pairs may belong to more than one category:

   Self-circles: The two reads of the pair belong to the same restriction fragment and both point outwards.

   Dangling-ends: The two reads of the pair belong to the same restriction fragment an both point inwards. Dangling-ends are useful because the allow to infer the mean insert size of the library.

   Error: The two reads of the pair belong to the same restriction fragment an both point in the same direction.

   Extra dangling-end: The two reads belong to different restriction fragments but they are close enough and pointing toward each other so the molecule likely comes from an incomplete digestion.

   Too close from restriction end: The start of one of the reads is closer than 5bp to a restriction end.

   Too short: One of the reads is in a fragment that is smaller than 100bp.

   Too large: One of the reads is in a fragment that is bigger than 100kb, that likely correspond to a repetitive element.

   Over-represented: One of the reads is in a fragment that is abnormally enriched with respect to the rest of the fragments in a particular library. This filter is not used in HiChIP experiment since this unequal representation is expected due to the enrichment of particular loci with the chosen antibody.

   Duplicated: Duplicated pairs coming from PCR artifacts. It is not expected to find several pairs composed by exactly the same reads.

   Random breaks: One of the reads is too far from a restriction end, above the maximum insert size calculated from dangling-ends.

   Reads that do not belong to any of these categories or just belong to the over-represented are kept for further processing.
3. *Contact matrix calculation and visualization*:
   Raw contact matrices are then calculated and stored in dense format from filtered reads

at 10kb resolution for further processing using the TadBit function *hic_map*. In addition to that, two types of compressed multiresolution matrix files are also generated: *cooler* files for visualization in *HiGlass* (Kerpedjiev et al. 2018) and *hic* files for visualization in *Juicebox* (Durand et al. 2016). Details and code on how to generate this files are also available in the gitlab repository (acemelthesis/HiChIP/1-Raw2Visual/raw2visual.md).

(b) Downstream analysis :

1. *Deriving ChIP-seq tracks from HiChIP dangling ends*:
   If we do not consider the 3D information coming from the pairing of the reads it is possible to obtain a linear profile that is equivalent to the one of a ChIP-seq experiment. This is useful to pinpoint those regions of the genome with enough coverage to extract useful contact information. We constructed those tracks based only on dangling-end reads to avoid digestion and ligation sources of bias as proposed in Mumbach et al. 2016. Briefly, we consider each dangling-end pair a ChIP-seq hit with the start and the end corresponding to the start and the end of the restriction fragment they belong to. These hits are represented in a bedfile that is then used to generate coverage profiles in bedgraph format using the *bedtools genomecov* tool. These bedgraph files are then further converted in bigwig in order to be visualized in the *USCS Genome Browser* using the *wigToBigWig* tool from *UCSC Kent Utils*. The profiles obtained are highly comparable to those seen in regular ChIP-seq experiments both qualitatively and quantitatively (Figure 3.5B, ChIP-seq experiments from Bogdanovic et al. 2012). In the case of the H3K4me3 experiments, the dangling-end derived bedfiles of the two replicates are also used to call peaks using *macs2* (Zhang et al. 2008) followed by *idr* correction (Li et al. 2011). The *macs2* parameters were the default ones with the following modifications: duplicates are kept since they are filtered before and the bedfiles contain the restriction fragment coordinates (−−keep-dup), we are extending 147 bps (−−nomodel, −−extsize 147) and we are using a loose p-value of 0.05 as recommended prior to *idr* correction (−p 0.05). Genome sizes used were 1.37 x 10^9 bps in the case of zebrafish experiments and 0.5 x 10^9 bps in the case of amphioxus experiments (specified with −g gsize). For the *idr* correction also default parameters are used and peaks with a p-value lower than 0.01 were selected and used in the prediction of RLs from the H3K4me3 HiChIP experiments in the next step. It is worth noting that there was an equivalent enrichment of traditionally derived vs HiChIP derived ChIP-seq signal around HiCHIP based H3K4me3 peaks (Figure 3.5B). The clustering was performed using *seqMiner* with default parameters (Ye et al. 2011). The scripts and the explanation needed to calculate these signal and the peak calling is available in the gitlab repository (acemelthesis/HiChIP/2-DownstreamAnalysis/chip_calling/chip_calling.md).
2. *HiChIP validation with 4C-seq*:
   In Figure4.10B there is a heatmap comparing the 4C-seq and the HiChIP data around 47 developmental gene promoters at 10kb resolution in zebrafish. Those genes are active according to the H3K4me3 occupancy, and therefore the HiChIP is informative about the 3D architecture around. In addition, they are genes with two available 4C-seq replicates at 24hpf. Briefly, the contacts in a 2Mb window around the promoters are binned in 10kb intervals and stacked in a matrix (four matrices in total, the 2 HiChIP and the 2 4C-seq

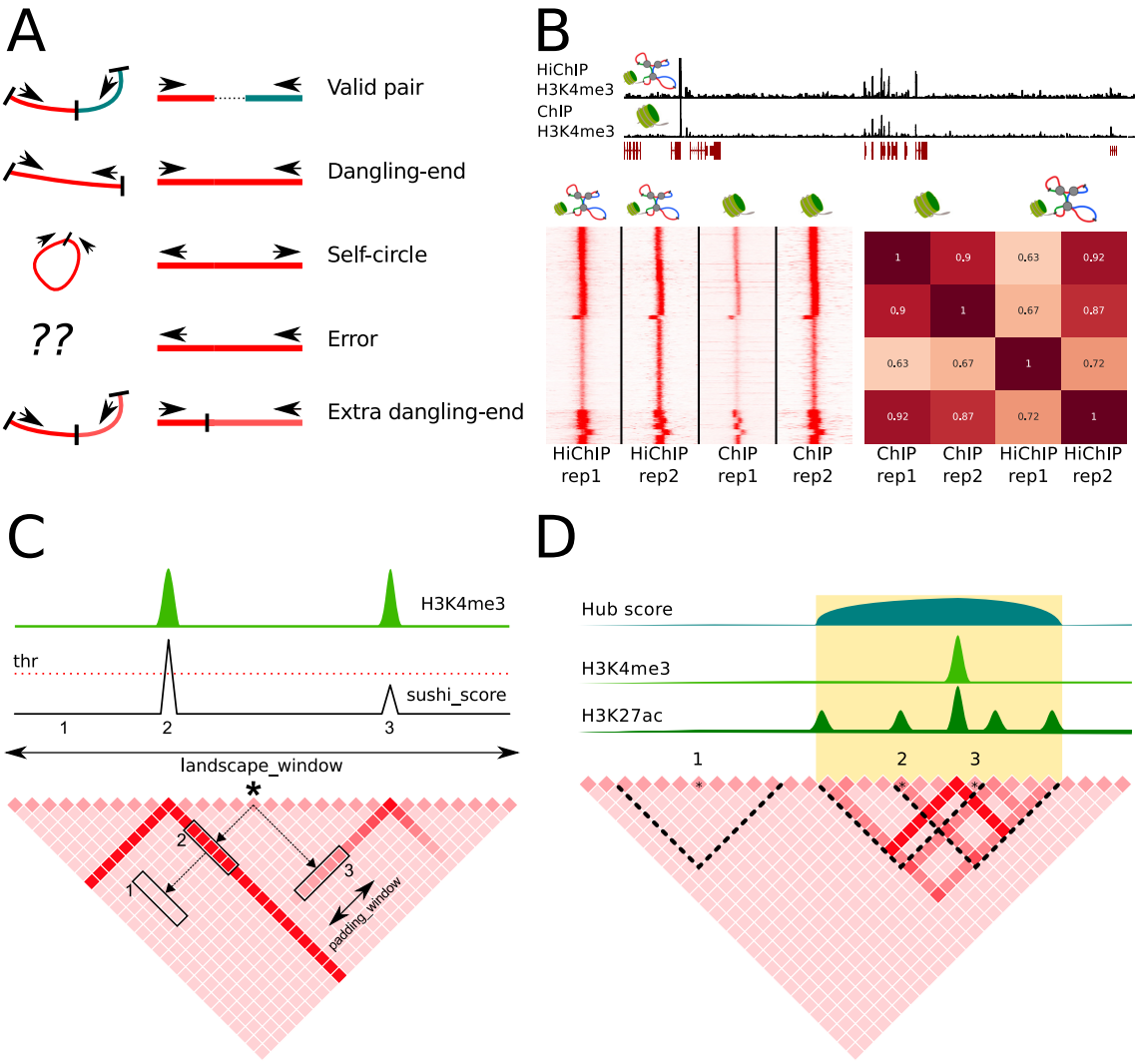


Figure 3.5: HiChIP analysis. (A) Scheme with some the ligation product (left) and mapping outcome (rigth) of some of the typical read pair types found after a HiC or HiChIP experiment. (B) Comparison between ChIP-seq and HiChIP enrichments. ChIP-seq like derived tracks are very similar to regular ChIP-seq tracks as seen in the tracks above. Bottom left there is a clustering with the signal of both kind of ChIP like experiments around peaks called in the HiChIP derived signal. The correlations of the signal matrices are shown in the bottom right part of the panel. (C) Scheme of the rationale of the *Sushibox* algorithm. A sushi_score is calculated for each pair of genomic loci in a landscape_window. In the example, the sushi_score of the loci 1,2 and 3 with respect to the reference loci (asterisk) is the sum of the signal of the squared bins (the bin connecting them plus a padding_window). The reference bin is assigned to the promoter 2 because is the only one above the threshold. (D) Scheme of the rationale behind the enhancer hub detection algorithm. A sliding triangle is moved along the chromosomes (discontinuous triangles). The sum of the signal inside those triangles correspond to the hub score. Enhancer-promoter hubs will have a high hub score.

replicates). The matrices are then normalized by row and plotted as heatmaps. Pearson and Spearmann correlations shown in Figure4.10C are pair-wise correlations of each of the normalized matrices. The code generating these analysis is in the gitlab repository (acemelthesis/HiChIP/2-DownstreamAnalysis/fourC_validation/HIChIPvs4C.ipynb).

3. *RL prediction from H3K4me3 HiChIP data (sushibox algorithm)*:
   The *Sushibox* algorithm was designed in order to predict RLs out of H3K4me3 HiChIP

experiments (see Figure 3.5C). Briefly, the algorithm scan each position in the contact matrix trying to find the closest promoters in 3D in a given *landscape_window* (by default 1Mb around each locus). In order to do that it calculates a *sushibox score* between the current position and the rest of loci in the *landscape_window*. This score corresponds to the sum of the signal of the bin connecting each pair of loci plus the immediate upstream and downstream neighboring bins according to the *padding_window* parameter (the default is five bins in each direction, corresponding to 50kb in 10kb resolution contact matrices). A threshold is then set according to the maximum *sushibox score* in the window (the default is 50% of the max value). The current position is then associated with all those bins above the threshold in both replicate experiments. Then an additional filter is used, and each position can only be associated with putative promoters predicted according to the H3K4me3 peak calling performed in the previous step. Finally, RLs are calculated by merging the genomic positions of the bins that were associated to each of the active promoters. *Sushibox* code and the explanation on how to use it is available in the gitlab repository (acemelthesis/HiChIP/2-DownstreamAnalysis/sushi_landscapes/sushi_usage.md).

4. *Analysis of the effect of WGDs in RL size*:
The analysis needed to produce the Figure 4.11 of the section 4.2.4 was performed using Python2 and the code is available in the gitlab repository (acemelthesis/HiChIP/2-DownstreamAnalysis/landscapes_wgds/RLsize_vs_WGDs.ipynb). Briefly, RL sizes were calculated from the *Sushibox* predicted RL boundaries both in zebrafish and in amphioxus by substracting the end from the start coordinates. With this information, the plot in Figure 4.11A was produced. Then, an orthology pre calculated with OMA (Altenhoff et al. 2018, used previously by Marlétaz et al. 2018) was used to keep a fraction of ohnolog genes: those paralogs that are retained in vertebrates in either a 1:1, 1:2, 1:3 or 1:4 relationship with respect to the number of copies in amphioxus. From those, only the ohnolog families in which all the members had predicted RLs by *Sushibox* in both species were kept. These families were further classified in developmental and housekeeping with a prediction used pereviously again by Marlétaz et al. 2018. RL sizes stratified both by species and by this developmental/housekeeping condition were plotted in Figure 4.11C. The RL sizes of the different members of the ohnolog families were also plotted stratified by the number of paralogs retained in zebrafish (Figure 4.11D). The number of developmental vs housekeeping genes in the different families depending on the number of vertebrate paralogs retained is also plotted in Figure 4.11E. Then, the RL sizes were plotted stratifed again by the number of ohnologs retained in zebrafish but ranking the sizes of zebrafish paralogs (Figure 4.7F, left). The number of ATAC-seq peaks inside each of the RLs from both zebrafish and amphioxus embryos of equivalent stages were quantified using *bedtools intersect* $-c$ (ATAC-seq experiments published in Marlétaz et al. 2018). With this, an equivalent plot to the previous but with the number of putative regulatory elements was done and shown in the right part of Figure 4.11F. Finally, the genes whose RLs experienced significant growth in the vertebrate lineage were singled out: specifically those that were bigger than the amphioxus copy and bigger than at least two of the zebrafish copies. Using the *ZEOGs* tool (Prykhozhij, Marsico, and Meijsing 2013), several expression domains enriched for genes with these

expanded zebrafish RLs were discovered and are shown in Figure 4.11G.

5. *RLs relationships with syntenic pairs*:
   The plots in Figure 4.12A and the results of the second part of the section 4.2.4 combine the information from the *Sushibox* predicted RLs and the age of syntenic pairs inferred in 3.3.2. Basically, the proportion of syntenic pairs included within the same RL depending on the age of the pair was calculated, using the *bedtools intersect* tool. Then, χ^2 statistics were applied to verify the influence of the age of the pairs. In addition, the collection of syntenic pairs that are vertebrate novelties and also share RLs in zebrafish according to the *Sushibox* predictions were extracted. The code to perform this is available in the gitlab repository (acemelthesis/HiChIP/2-DownstreamAnalysis/landscapes_synteny/RL-Synteny.ipynb).

6. *Enhancer-promoter hub predictions from H3K27ac data*:
   Seemingly simultaneus associations between promoters and several enhancers were predicted using H3K27ac HiChIP data for the results shown in Figure 4.13 and the tables 4.3 and 4.4. For that purpose, H3K27ac peaks were called from the HiChIP derived ChIP signal as explained in the step 1 of this section, using a threshold of 0.1 for the idr corrected p-values. Those peaks were then classified between enhancers and promoters according to their distance to the closest TSS (promoters were those H3K27ac peaks closer than 10kb to the nearest TSS). Then, we used the raw matrices at 10kb resolution to calculate a slightly different version of an insulation score to predict enhancer hubs (see Figure 3.5D) following the rationale that those loci resemble isolated TADs. Insulation scores that were unusually high within a chromosome (above 3Q + 1.5 x IQR) were kept and adjacent loci over the threshold were merged together and considered enhancer hubs. Then, hubs without enhancers (withouth H3K27ac peaks far from promoters) were filtered out. Finally, for each of the hubs, the candidate promoters were singled out. The total signal related to each of the bins in the hub was added together and again unusually high values were kept. If those points in the hub were closer than 30kb to a promoter (H3K27ac peak near a TSS) those promoters are kept as the plausible responsive genes to the enhancers in the hub. Overlapping hubs from the two replicates were merged together. Biological process Gene Ontology Enrichment was performed using the DAVID platform (Huang, Sherman, and Lempicki 2009). The code for this analysis and the instructions on how to use it is available in the gitlab repository (acemelthesis/HiChIP/2-DownstreamAnalysis/ehubs/ehub.md).

7. *A/B compartment prediction from HiChIP and HiC data*:
   The compartment prediction both from HiC and HiChIP experiments was performed using Juicer Tools (Durand et al. 2016). This is shown in Figure 4.14. Pearson transformed matrices at 100kb resolution were calculated using *juicertools pearson* and first eigenvectors using *juicertools eigenvector*. This was done for the chromosome 25 in zebrafish HiC (Kaaij et al. 2018) and H3K27ac and H3K27me3 HiChIP experiments; for the chromosome 3L in *Drosophila* in equivalent experiments (Rowley et al. 2017) and only in the two kind of HiChIP experiments for the Sc0000005 in amphioxus. H3K27ac and H3K27me3 HiChIP derived signals were binarized in 100kb chunks to make them comparable with the compartment signal and were plotted along with the pearson matrices. Genomic bins with an eigenvector value of different sign were considered to be belonging to different compartments. The compartment enriched

in H3K27ac signal was considered to be compartment A. HiC experiments were obtained from GEO repositories and analyzed as explained in section 3.2.3. The code and the details are again available in the gitlab repository (acemelthesis/HiChIP/2-DownstreamAnalysis/compartments/compartments.md).

3.2.3

HiC data analysis

HiC experiments from zebrafish (Kaaij et al. 2018) and *Drosophila* (Rowley et al. 2017) were reanalyzed in an almost equivalent manner to the HiChIP experiments as explained in the *From raw fastq files to contact visualization* part of section 3.2.2. The only difference was that over-represented reads were filtered out in this case. HiC experiments were used only for visualization and for calculating compartments as explained in the last step of the previous section (3.5). The modified version of the pipeline that analyze HiC experiments can be found in the gitlab repository (acemelthesis/HiC/hic.py).

3.3

Analysis of microsynteny

Microsynteny analysis of the neighboring regions of the different Hox loci were key in order to draw many of the conclussions of the section 4.1 of the results, including the ones in Figure 4.2 and Figure 4.6. In this case the syntenic comparisons were done manually following the criteria stated in Acemel et al. 2016. Briefly, different genomes were browsed using the NCBI, UCSC and Ensembl Metazoa platforms. The species studied and their assemblies included: elephant shark (*Callorhinchus milii*, 6.1.3), mouse (*Mus musculus*, mm10), the limpet *Lottia gigantea* (v1.0), the flatworm *Trichoplax adhaerens* (v1.0), the centipede *Strigamia maritima* (Smar1.0), the hemichordate *Saccoglossus kowalewskii* (Skow_1.1) and the starfish *Acanthaster plancii* (v1.0). TBLASTN and BLASTP tools were used when annotations were either absent (like in the case of the starfish) or not complete. In order to find the pseudogenized exons of *Jazf2* in the mouse HoxD vincinity, the VISTA tool (Frazer et al. 2004) was used (see Figure 4.2C). Elephant shark was used as a reference and LAGAN as the alignemnt program with the following parameters: 100-bp window and 65% identity in 70 bp. In the following sections, however, we will concentrate in how synteny was calculated automatically for the results in the section 4.2.

3.3.1

Prediction of microsyntenic blocks between zebrafish and mouse

For the results in the section 4.2.2, and particularly in the Figure 4.8B, an automatic prediction of sytenic blocks between zebrafish and mouse was used. Precomputed chain alignments between the mouse *mm9* and the zebrafish *danRer10* assemblies were downloaded from the UCSC genome browser and processed. The zebrafish genome was the target genome and the mouse the query. Briefly, the biggest chains overlapping zebrafish genes with available 4C-seq experiments

were selected as seed blocks. In addition, the rest of the chains inside the automatically predicted RLs from the 4C-seq experiments (see section 3.1.2) were also kept. The seed blocks could be later updated by extending their upstream and downstream limits with some of these other chains. The criteria used was that these chains and the seed block must be closer than 2Mb in the mouse reference. The Python code to perform these analysis are available in the gitlab repository (acemelthesis/synteny/syntenic_blocks/Chains-in-RLs.ipynb). In order to generate the Figure 4.8B these syntenic blocks were analyzed together with the RL boundaries predicted by 4C-seq as described in the part 5 of the section 3.1.2.

3.3.2

Microsyntenic pair analysis in the deuterostome lineage

Syntenic pairs analysis following the rationale of those performed in Irimia et al. 2012 were used to estimate the age of every pair of genes in both zebrafish and amphioxus genomes. We used the annotations of nine representative genomes to perform this analysis: two amphioxus species (*Branchiostoma lanceolatum, Bl71nemr* assembly; *Branchiostoma belcheri, Haploidv18h27* assembly), two non chordate deuterostomes (the hemichordate *Saccoglossus kowalewskii, Skow_1.1* assembly; the echinoderm *Strongylocentrotus purpuratus, strPur4* assembly) and five vertebrates: zebrafish (*Danio rerio, danRer10* assembly), medaka (*Oryzia latipes, oryLat2* assembly), chicken (*Gallus gallus, galGal5* assembly), mouse (*Mus musculus, mm10* assembly) and human (*Homo sapiens, hg38* assembly). We also took advantage of the OMA orthology used previously for the analysis of the ohnologs (see section 3.2.2, downstream analysis part 3). Briefly, first genes without orthologs in any of the other species were filtered out from the annotations. Then, syntenic pairs are searched using as the reference each one of the species and as target the remaining eight. In order to do so, we sought for the orthologous genes in the target annotation of each pair of contiguous genes in the reference annotation. Then we checked how many other genes we could find in the target annotation in between our pair of contiguous genes in the reference annotation (using $bedtools\ intersect - c$). If they were less than four, we kept the pair as syntenic. Then, we further filtered out pairs containing genes that were not ohnologs (i.e. that were not retained in 1:1, 1:2, 1:3 or 1:4 paralog patterns when comparing mouse and amphioxus) to exclude hotspots for tandem duplications. We repeated this using as reference each of the nine genomes and merged the data removing the duplicated information. At this point it was possible to roughly date the different pairs of genes of zebrafish and amphioxus. We considered a pair to be amphioxus or zebrafish specific if it could be only found in either of them. In the case of zebrafish, if the pair was also present in the other four vertebrate genomes, but not anywhere else, it was dated to the LCA of vertebrates. If the pair was also found in one of the two amphioxus genomes but not anywhere else, it was dated to the LCA of chordates. Finally, if the pair was also found either in *Saccoglossus kowalewskii* or in *Strongylocentrotus purpuratus* it was dated to the LCA of deuterostomes (at least). Further details, the code and the instructions to reproduce this analysis is available in the gitlab repository (acemelthesis/synteny/syntenic_pairs/date_pairs.md). The dating of the pairs is used in the HiChIP analysis part (see section 3.2.2, downstream analysis part 4; and the results are in section 4.2.4 including Figure 4.12 p.103).

3.4

Transgenic reporter assays for enhancer detection

Transgenic reporter assays in zebrafish were performed in order to assess and characterize the enhancer activity of several putative regulatory regions from amphioxus identified using ATAC-seq. The results are presented in Figure 4.5C and in Figure 4.8C. Briefly, the PCR amplified sequences of the different putative enhancers were subcloned in PCR8/GW/TOPO vectors (Thermo Fisher, K250020). They were then shuttled to the enhancer detection vector using the Gateway technology (Thermo Fischer, 11791100). The sequence of the primers are shown in Table 3.1. This enhancer detection vector contain the *gata2* minimal promoter, GFP, an strong midbrain enhancer (z48, used as an internal control and reported in Calle-Mustienes et al. 2005) and the Tol2 transposase recognition sequences (Kawakami 2004). Then, one-cell stage fertilized embryos were microinyected with a solution of 25ng/μL of the mRNA of the Tol2 transposase, 20ng/μL of the purified reporter vectors and 0.05% of phenol red. Three or more independent and stable transgenic lines were generated for each of the constructs, and all of them displayed an equivalent GFP distrubution during embryogenesis.

EndE_F	GCGATTAAGATGGAAGTAGGATGC
EndE_R	TTGTTGATGACCCTTATGCACAC
1655_F	TTGTTGATGACCCTTATGCACAC
1655_R	CGGGTGCCAGGTTTATTCGA
1739_F	GCTGAGATTTCCAAACAACCACA
1739_R	GGGACACGGAGGTTGATAAGT
1784_F	CCACCCGGAAATCTTTGTCC
1784_R	TATGCGCTCTGAGATGACGG
1801_F	TTCCGCATGCCTTACACACA
1801_R	CCCGCGATATAAAGCCCAGT
2473_F	AGTCACCCGTTAGATTCCCT
2473_R	ACCATACGCTGCTTATCCATGA

Table 3.1: Primers for the amplification of the putative amphioxus enhancers for the reporter assays.

Chapter 4

Results

4.1

The evolution of the 3D architecture of the vertebrate HoxD locus

4.1.1

The topology in two TADs of the HoxD RL was likely the ancestral configuration before vertebrate WGDs

In order to understand how the complex 3D topology surrounding the HoxD loci originated we decided to infer which configuration was more likely present in the last common ancestor of all extant vertebrates. As a reminder, the HoxD genes are located precisely over a TAD boundary, benefiting from regulatory information coming from both the anterior and the posterior TAD (Figure 4.1A) During limb development, genes located towards the anterior end of the cluster (*HoxD1* to *HoxD7*) tend to interact more with enhancers from the anterior TAD while posterior genes (*HoxD12* and *HoxD13*) tend to interact more with the posterior TAD. Intermediate genes (*HoxD8* to *HoxD11*) are able to switch from reading enhancers from the anterior TAD to the posterior depending on the cell population during limb development. Therefore, the topology in two TADs is critical for the proper establishment of this regulatory mechanism. Equivalent dynamics have been also reported for the mammalian HoxA cluster that is also involved in the developing limbs (Figure 4.1A).

In the introduction we anticipated that the origin of vertebrates came coupled to two events of WGDs that gave rise to 4 copies for each loci of the last preduplicative ancestor. Many of these duplicated genes were secondarily lost, but all four Hox clusters have been maintained in mammals. Strikingly, up to eight of them are present in teleost fishes due to an extra round of WGD that was specific to this group (Figure 4.1B). Taking this into account we explored available HiC data from human and mouse around the four mammalian Hox clusters. In humans, we checked HiC data from the cell lines GM12878, IMR90, HMEC, NHEK, K562, HUVEC, HeLa and KBM7. In mouse we explored the CH12-LX and the J1 mESC cell lines plus HiC experiments done in cortex (Dixon

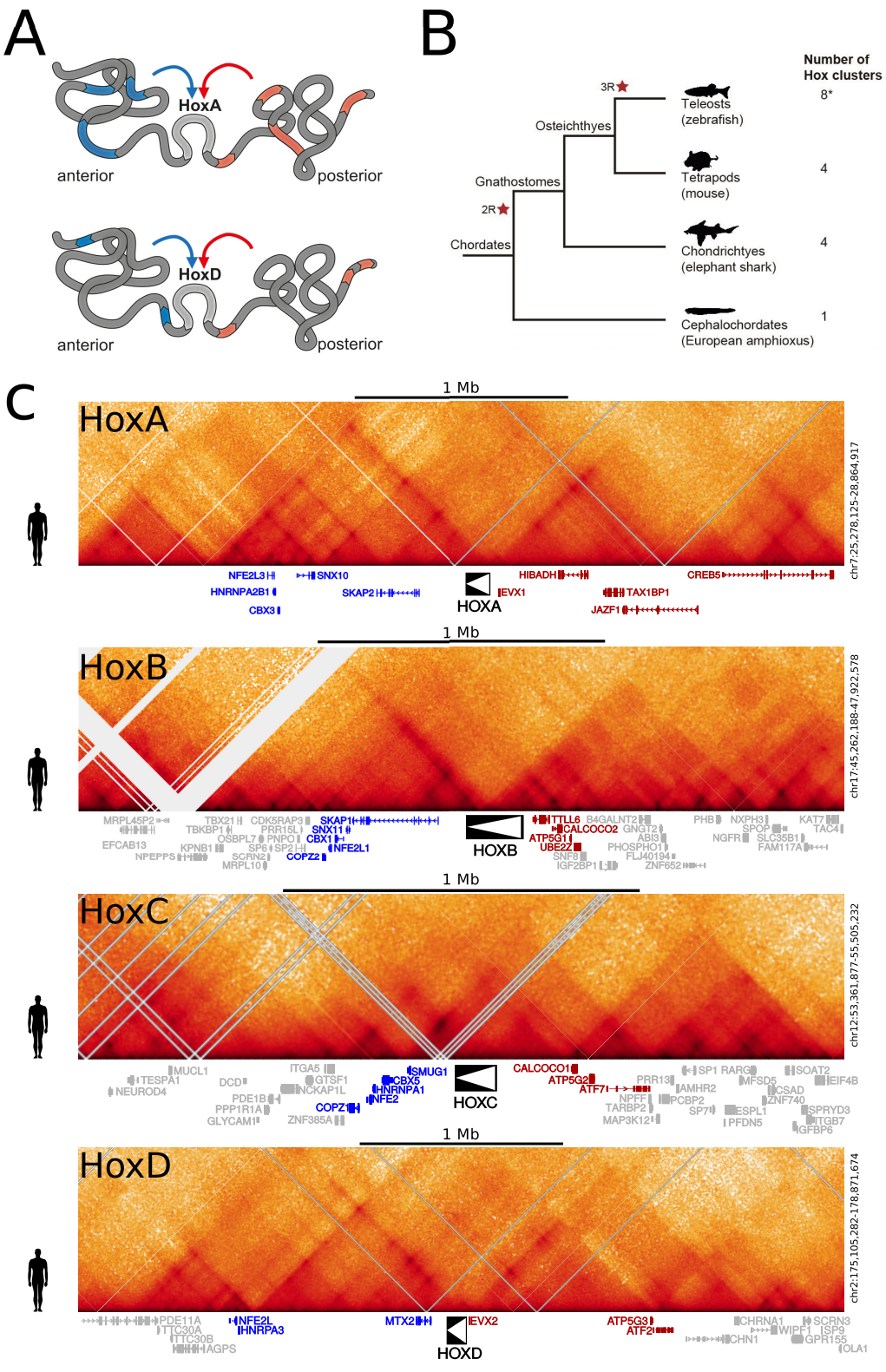


Figure 4.1: Split configuration of the Hox locus in the vertebrate ancestor. (A) Both the murine HoxA and HoxD clusters have been shown to be located at the boundary between two TADs benefiting from regulatory information present in either of them. (B) Hox genes tend to be retained after WGD events, with four of them present in most vertebrates and up to eight in teleost fishes. (C) HiC matrices around human Hox clusters in the human cell line GM12878 at 10kb resolution (from Rao et al. 2014). Heatmaps were generated with *Higlass* (Kerpedjiev et al. 2018). The bipartite configuration in two TADs is evident at least in the HoxA, HoxB and HoxD clusters.

et al. 2012) and limb bud tissues (Rodríguez-Carballo et al. 2017). As expected, in both species the 3D configuration in two TADs was readily visible in the HoxA and HoxD clusters, but also in the HoxB (Figure 4.1C, for simplicity only data from the human GM12878 is shown). The TAD boundary location at the middle of the Hox cluster was remarkably stable across different cell lines and tissues in both species even though Hox genes are not expressed in some of them. In contrast, the structure of the HoxC cluster seemed a little more disorganized. Nevertheless it could still be argued that there is a slight contact insulation between the anterior and the posterior end of the HoxC cluster. In any case, since the bipartite topology in two TADs is present in at least three out of the four clusters, the most parsimonious scenario is that this topology was already present in the only Hox cluster of the last common ancestor of vertebrates before the two WGD events.

To explore this further we decided to study the synteny of the anterior and the posterior regions flanking the four vertebrate Hox clusters, together with the regions flanking the only cluster present in the cephalochordate amphioxus (*Branchiostoma lanceolatum*) as an outgroup. A remarkable conservation of microsynteny was found in both the anterior and the posterior flanking regions between the four different paralog clusters in vertebrates (Figure 4.2A). Strikingly, the synteny of the anterior flanking region was also shared with the anterior flanking region of the amphioxus Hox cluster. In addition, *Evx* and *Lunapark* genes, which are immediately adjacent to the posterior end of HoxA and the HoxD clusters, were also found at the posterior end of the amphioxus Hox cluster. However, the remaining part of the posterior flanking region in amphioxus is not syntenic to those of the vertebrate clusters.

It is important to note that these syntenic equivalences could be assessed even though most of the genes that are considered to be in synteny are not conserved around all four vertebrate clusters due to events of differential gene loss. These events are pervasive after WGDs and indeed only *Nfe2* could be found both in the anterior regions of all four vertebrate clusters and in amphioxus. Five other genes (*Hnrpa, Cbx1-3-5, Calcoco, Atp5g* and *Creb*) were found in at least three of the four vertebrate clusters. In contrast, others were only found flanking two vertebrate clusters (*Copz, Skap*) or even only one (*Smug1, Mtx2, Lunapark*). However, they are found in equivalent positions and orientations in amphioxus, suggesting that they were also present in the preduplicative ancestor. Finally, we have the case of the *Hibadh/Jazf/Ube2z* block that most times is only found in the vertebrate specific posterior region of the HoxA cluster. Therefore it was challenging to trace its origin. Luckily, a paralog of *Jazf* is also found in an equivalent region posterior to the HoxD cluster in elephant shark (Figure 4.2B). Furthermore, inside the introns of this elephant shark paralog there are two conserved enhancer sequences (I.1 and I.2) that are also present in the flanking region posterior to the mouse HoxD cluster. This indicates that both genomic regions are likely equivalent and that the exons of the *Jazf* paralog lying close to the HoxD cluster were erased in the mouse ancestor while the enhancers were maintained. In fact, it is still possible to find a pseudogenized version of a *Jazf* exon near to the I.1 enhancer in mouse (Figure 4.2C). Similarly, *Hibadh* carries in one of its introns the conserved enhancer CsBA, that drives the expression of HoxA genes in the limb buds. No trace of any *Hibadh* exon is found in the equivalent intergenic region around the HoxD cluster, but there is an enhancer that is homologous to the CsBA, called CsBB (Figure 4.2B). Since *Hibadh, Jazf* and *Ube2z* form a conserved microsyntenic block present also outside vertebrates, it is reasonable to think that the whole block was present in the posterior flanking regions of the preduplicative ancestor.

To sum up, the conservation of microsynteny between the regions flanking the four vertebrate Hox clusters further suggests that there was distal regulatory information contained within those

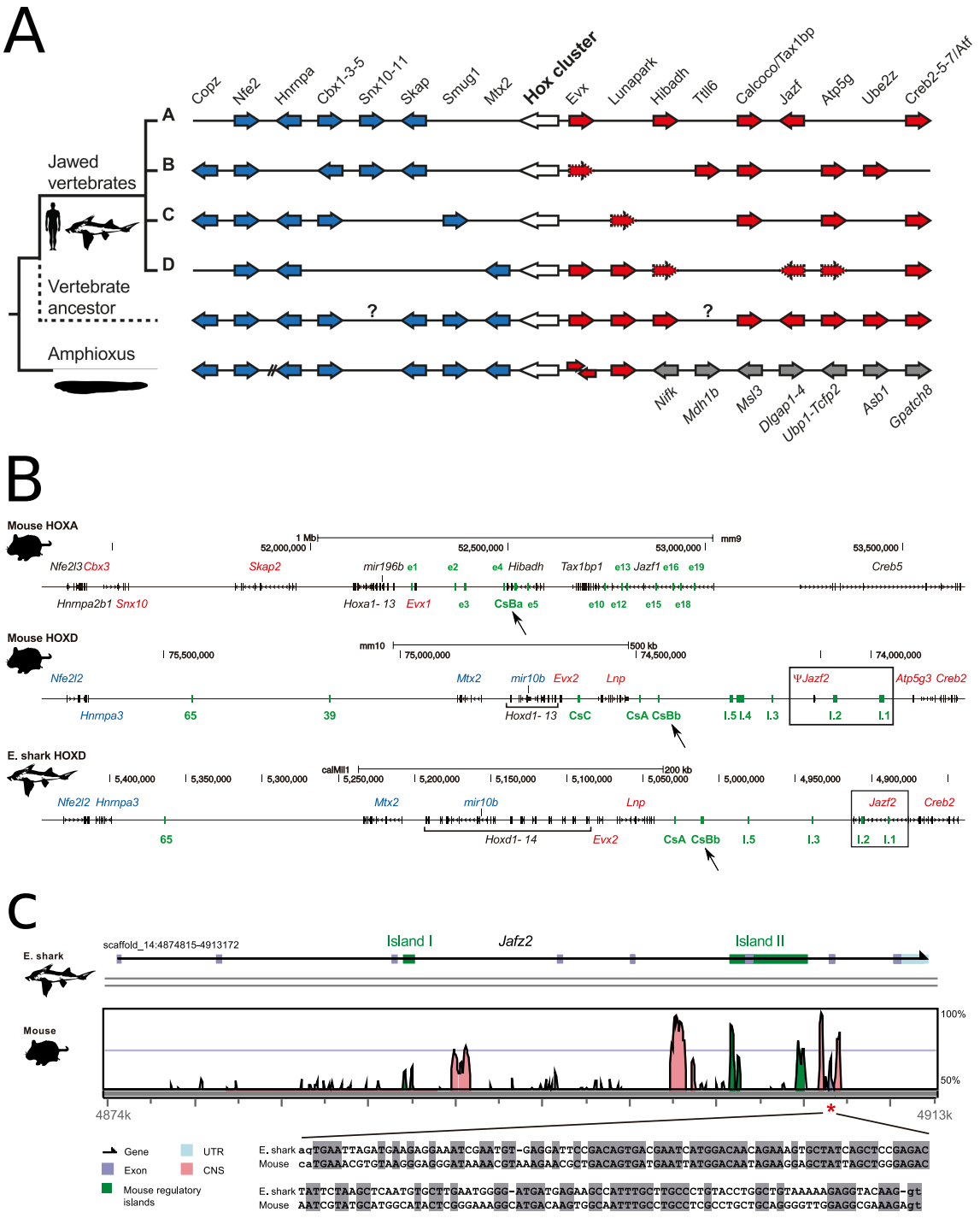


Figure 4.2: Shared synteny around vertebrate Hox loci. (A) Genes surrounding all four paralog Hox clusters are largely equivalent although affected by differential gene loss processes. Syntenic genes from the anterior TAD are colored blue while syntenic genes from the posterior TAD are colored red. The anterior neighboring region of amphioxus is also homologous to the anterior vertebrate neighboring region. (B) Differential gene loss in action: the posterior TAD of elephant shark HoxD locus contains a *Jazf2* copy (rectangle), while in mouse we only found *Jazf1* associated to the HoxA cluster. Arrows show how a conserved limb enhancer (CsB) can be found both in the HoxA and the HoxD clusters in mouse and elephant shark. (C) A pseudogenized *Jazf2* is still recognizable associated to the murine HoxD locus through VISTA alignment (also highlighted with a rectangle in (B), Frazer et al. 2004, Material and Methods section 3.3).

regions already in the only cluster of the last ancestor of all vertebrates before the two WGD events. Taking this together with the 3D configuration in two TADs of three out of the four clusters both in mouse and in human, it seems likely that this topology was the ancestral for the vertebrate lineage. Interestingly, the synteny constraints of the anterior flanking region are shared between vertebrates and amphioxus suggesting that these region might be already wired to the Hox regulation in the last common ancestor of chordates. To further test this we decided to explore the 3D configuration of the amphioxus Hox cluster and compare it with the topology in two TADs found in vertebrates.

4.1.2

4C-seq experiments coupled to 3D modelling revealed no bipartite Hox regulation in amphioxus

With the purpose of characterizing the 3D organization of the chromatin around the amphioxus Hox cluster we designed an array of 14 4C-seq experiments with baits covering a region of 1.6 Mb. The bait selection included 7 Hox genes, namely *Hox2, Hox5, Hox6, Hox7, Hox9, Hox11, Hox13* and *Hox15*. In addition, we included three baits from both the anterior and the posterior flanking regions. At the anterior end we chose two genes that were contained within the genomic region that was syntenic with vertebrates (*Hnrpa* and *Mtx2*) and also *Meox*, that is beyond this region. Posterior to the Hox cluster we included the only two syntenic genes (*EvxA* and *Lnp*) plus *Gpatch8*, that is located further downstream. Besides, we also decided to profile the zebrafish HoxDa locus using an equivalent approach, so that it is possible to compare the topology of the amphioxus Hox cluster with the topology of the vertebrate HoxD cluster in a fair manner. For that we used nine extra 4C-seq experiments in zebrafish with the baits spanning a region of 0.7 Mb approximately including 4 Hox genes (*hoxd4a, hoxd10a, hoxd11a* and *hoxd13a*) and five extra genes, two in the syntenic part of the anterior flanking region (*nfe2l2a* and *hnrpa3*) and three in the posterior (*evx2, lnpa* and *atp5g3a*). We performed duplicated experiments in 24hpf zebrafish embryos and in three different stages of amphioxus development: 8hpf, 15hpf and 36hpf. Statistically significant interacting regions were defined for each experiment and the interaction events inside those regions were quantified (see Materials and Methods, 3.1.2, p.56). In order to interpret these numbers we followed the rationale that the 4C-seq of a promoter located in the middle of a TAD will likely have no left/right bias in its interaction pattern. Meanwhile, a promoter with a TAD boundary closer to its left will have its interaction pattern biased to the right, since contacts towards the left are hindered by the boundary. The opposite would happen to promoters with TAD boundaries closer to their right.

First of all, the bipartite configuration in two TADs of the zebrafish HoxDa cluster was obvious upon visual inspection. Despite being separated only by 38kb, the promoters of *hoxd4a* and *hoxd13a* located at both ends of the cluster displayed remarkably opposite interaction patterns (Figure 4.3A). Indeed, 83% of the contacts of *hoxd4a* were directed towards the anterior flanking region (to the left) while the 79% of the contacts of *hoxd13a* were directed towards the posterior one (right). Intervening Hox genes showed intermediate contact patterns but always biased towards the anterior end (71% and 67% showed by *hoxd10a* and *hoxd11a* respectively). This sharp transition is compatible with the presence of a strong boundary bisecting the HoxDa cluster in two TADs, similarly to what is seen in mammals. Furthermore, the two syntenic promoters of the anterior flanking region (*nfe2l2a* and *hnrnpa3*) showed almost no interactions with the posterior flanking

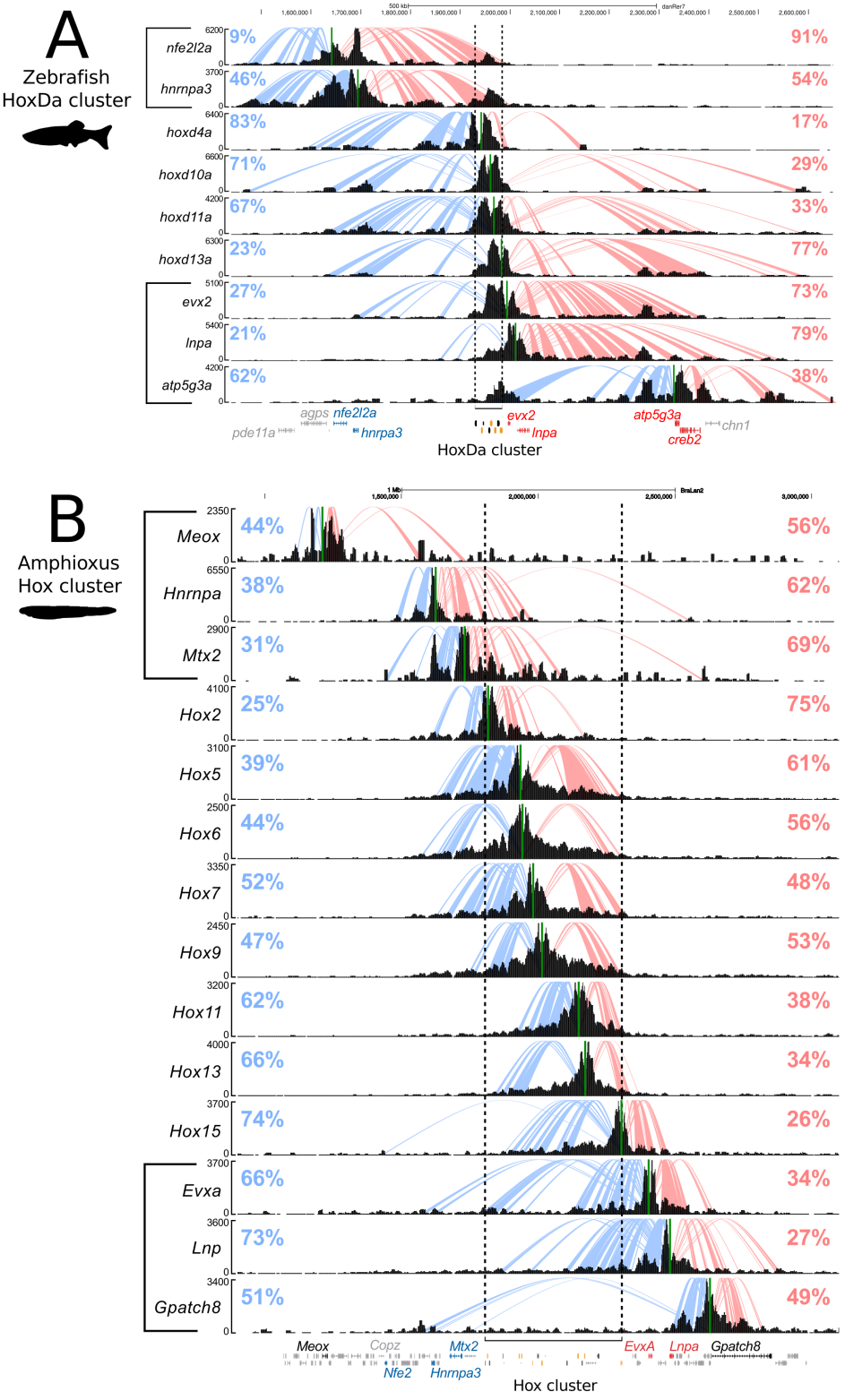


Figure 4.3: 4C-seq experiments around the zebrafish HoxDa (A) and the amphioxus Hox (B) loci. Experiments are performed in whole embryos of 24hpf in zebrafish and 15hpf in amphioxus. The arachnograms highlight the significant interactions with the bait. The percentages are the number of reads at each side of the viewpoint inside significant interactions. Genes written blue and red are those of the anterior and posterior syntenic regions respectively. Dashed lines indicate the extension of the cluster.

regions (below 3%). Moreover, from the posterior side, *atp5g3a* showed no significant interactions with the anterior flanking region. Accordingly, *evx2* and *lnpa* interactions were also extremely biased towards the posterior side (73% and 79% respectively). These observations further indicate that the anterior flanking region is strongly insulated from the posterior one.

In contrast, such a transition point compatible with a TAD boundary could not be found inside the amphioxus Hox cluster in 15hpf embryos, despite the fact that the cluster occupies 400kb more than the HoxDa cluster of zebrafish (Figure 4.3B). First, we observed that from *Hnrnpa* to *Hox2* contacts are strongly biased towards the cluster, indicating the presence of a close boundary anteriorly. At the same time, *Meox* contacts showed little directionality bias and almost no overlap with *Hnrnpa*, suggesting that the insulating point is located between *Meox* and *Hnrnpa*. Besides, the promoters of the genes ranging from *Hox2* to *Hox9* interacted with both sides evenly, which suggests that the region does not contain a TAD boundary as observed in vertebrates. Finally, the promoters from *Hox11* to *Lnp* are biased anteriorly indicating the presence of a boundary close to the latter gene. This boundary is probably placed between *Lnp* and *Gpatch8* since the promoter of *Gpatch8* displays little directionality bias and their contacts have little overlap with those of *Lnp*, similarly to what was seen between *Meox* and *Hnrnpa*. In summary, the 4C-seq data suggest that the whole set of Hox genes seems to be embedded within a single domain together with the anterior syntenic region, *Evx* and *Lnp*. This contrasts with the vertebrate HoxD scenario, where a TAD boundary is located in between the cluster and posterior Hox genes contact enhancer sequences far beyond *Lnp*. In fact, arguably, *Evx* and *Lnp* promoters in amphioxus seem to be already insulated from the rest of the Hox TAD in a small sub-domain.

4C-seq data as is allowed us to confirm the bipartite configuration of the HoxDa cluster in zebrafish and also suggested the absence of this configuration in amphioxus. However, 4C-seq data only provide a limited set of pairwise distances, that are those that involve the selected viewpoints. Interactions between non-viewpoint loci are unknown, and we wondered if those could illuminate alternative topologies explaining the amphioxus 4C-seq profiles other than the presence of a single domain containing the anterior syntenic region, all Hox genes and arguably also *Evx* and *Lnp*. In order to calculate those distances we decided to apply integrative modeling approaches, inspired by studies using distance restraints to infer the architecture of protein complexes such as the exocyst (Picco et al. 2017). Hence, the chromatin of both loci was represented as a string of beads and next the interaction frequencies obtained from the 4C-seq experiments were used as distance restraints to position those beads. Once the models were generated, a whole set of distances between every loci in the region was available, and those distances may be represented either as 3D models or using heatmaps resembling those used for 5C or HiC experiments (Figure 4.4A). The two models largely confirmed what the 4C-seq data alone was suggesting. The zebrafish HoxDa cluster was located precisely at the hinge between two big interacting domains, while in amphioxus the entire cluster was placed inside a single domain together with the anterior syntenic region. In order to validate the modeling, several mouse loci with available 4C-seq information were also modeled. The resulting matrices were compared to those obtained by regular HiC methods and high correlations were obtained. In addition, the robustness of the method to the elimination of the information from several viewpoints was also tested. Some more details about the validation can be found in Material and Methods (section 3.1.3, p.60) and are also published elsewhere (Irastorza-Azcarate et al. 2018).

It is also important to note that the resulting 3D models are averages of the presumably dynamic chromatin interactions from millions of cells sampled from whole zebrafish and amphioxus embryos.

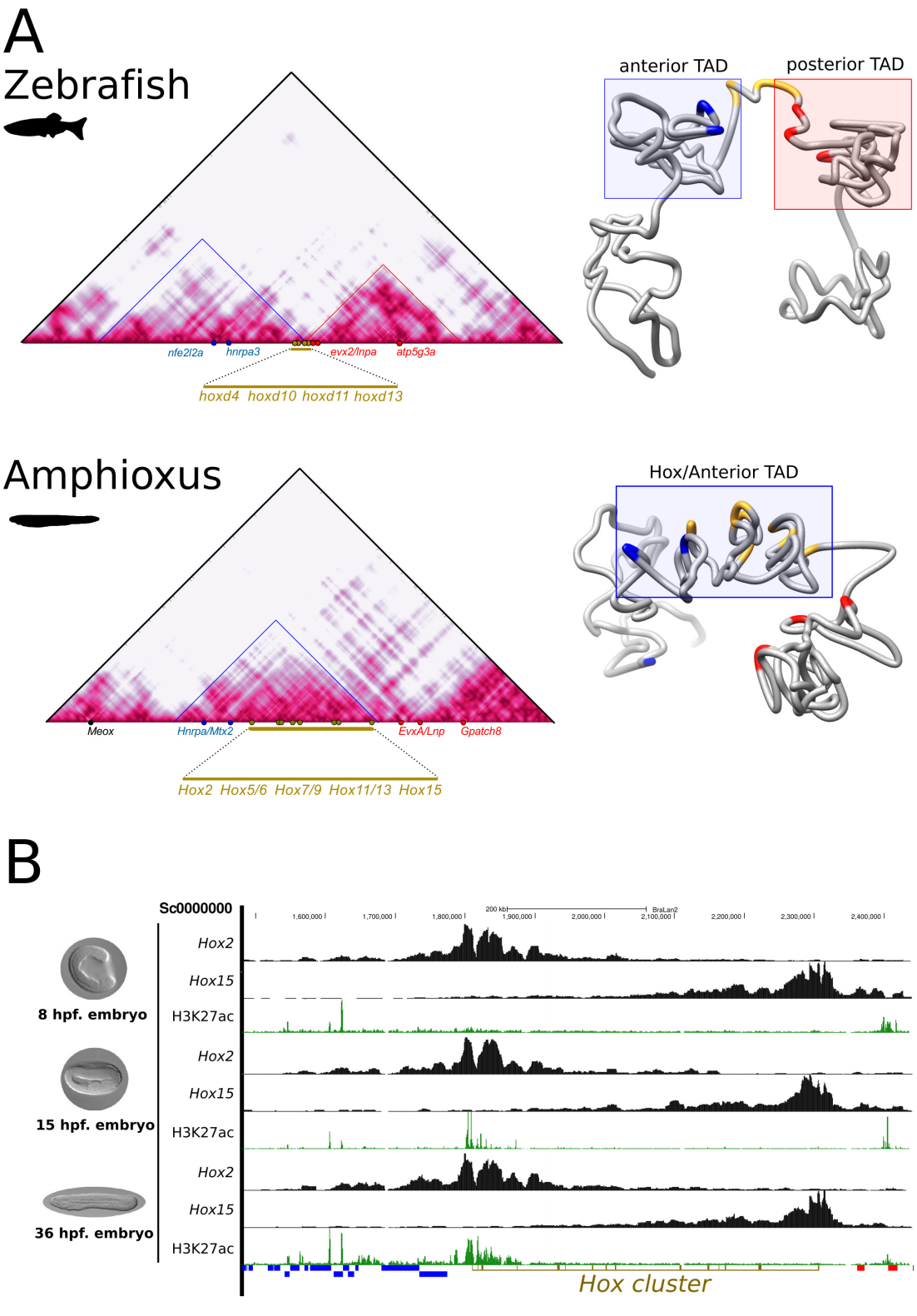


Figure 4.4: 3D model comparison. (A) Models (right) and v-HiC (left) showing the chromatin folding around the vertebrate HoxDa (top) and the amphioxus Hox (bottom) loci. (B) Contacts are almost identical in 4C-seqs from earlier and later amphioxus embryos (8hpf and 36hpf respectively) despite the progressive activation of Hox genes shown by the H3K27ac ChIP-seq tracks (Marlétaz et al. 2018)

.

Therefore, they probably will not represent precise 3D configurations of any particular cell, but, as we mentioned, an average profile informative of what regions promoters tend to interact with. We wondered how dynamic the topology of the amphioxus cluster was and therefore we compared our data from 15hpf embryos with the one generated in 8hpf and 36hpf embryos. We found a remarkable stability of the interactions between all three stages despite the progressive epigenetic and expression changes between them (Figure 4.4B).

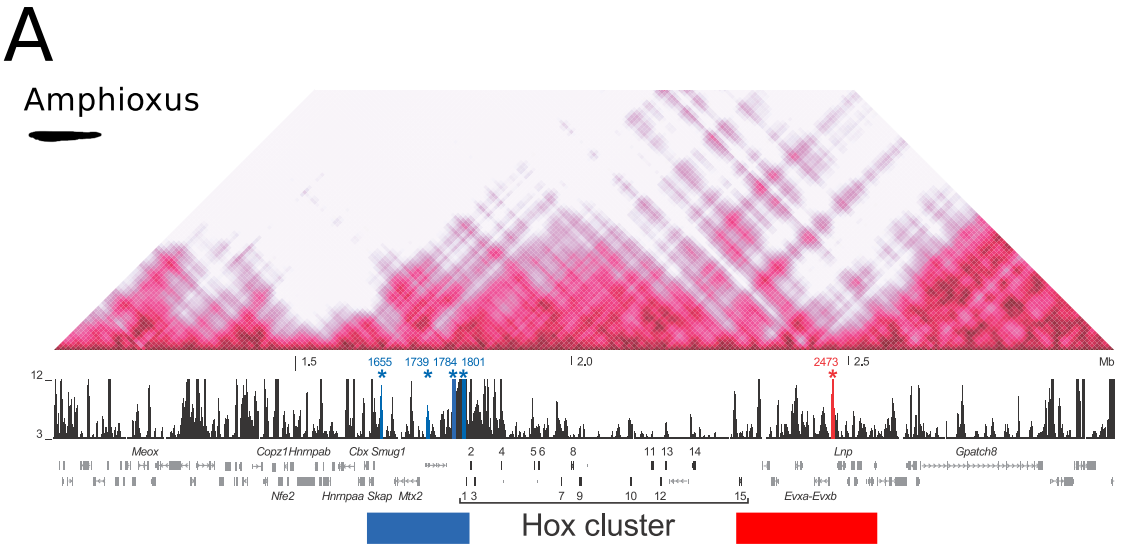
4.1.3

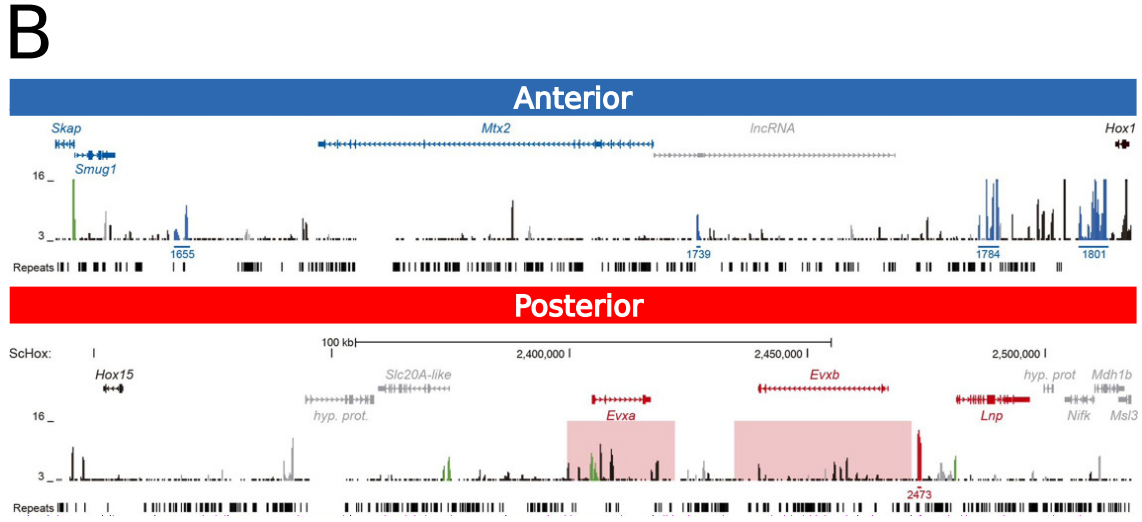
The syntenic region anterior to Hox1 is functionally wired to Hox regulation in amphioxus

We have already shown that the anterior region flanking the amphioxus Hox cluster is syntenic to the anterior region flanking the four Hox clusters of vertebrates. These anterior regions in vertebrates are populated with Hox specific enhancers. In addition, we have shown that the 3D topology of the Hox locus in amphioxus allows interactions between this genomic region and Hox promoters. We wondered if this anterior genomic region of amphioxus also contained regulatory sequences that were incorporated to control the expression of Hox genes. In order to do that we generated ATAC-seq experiments in 36hpf amphioxus embryos, allowing us to predict enhancers and promoters by detecting chromatin regions that are accessible to the TFs and the transcription machinery.

We manually inspected the ATAC-seq profiles in both the anterior and the posterior flanking regions and already inferred that both regions had a very different regulatory potential (Figure 4.5A-B). The anterior region was rich in open chromatin regions that were well separated from transcriptional start sites, potentially being distal enhancers. In stark contrast, only one putative enhancer not overlapping transcriptional start sites and repetitive elements could be identified at the posterior side. Therefore, tested the enhancer activity of four putative enhancers from the anterior flanking region, together with the only candidate from the posterior side (Figure 4.5C). We used a Z48-GFP-Tol2 system (Gehrke et al. 2014) and generated stable lines of transgenic zebrafish embryos carrying each one of the tested enhancers (see Material and Methods section 3.4, p.74). All four anterior enhancer candidates (named 1655, 1739, 1784 and 1801 and located 150kb, 66kb, 20kb and 3kb away from the *Hox1* promoter respectively) showed enhancer activity either in the spinal cord, in the hindbrain or in both. These expression domains were compatible with the reported endogenous expression of Hox genes in amphioxus but not with the expression domains of other neighboring genes. Interestingly, some of regulatory elements also displayed enhancer activity in vertebrate specific cell populations such as the neural crest and the otic and the olfactory placodes.

Conversely, the only putative enhancer found in the posterior side (named 2173 and located 165 kb away from *Hox15*) showed activity in some cells of the eye and in the neuroepithelium. This enhancer is located between *EvxA* and *Lnp*, and likely belong to the former since *EvxA* has been shown to be expressed in amphioxus neuroepithelial cells. These findings, together with the topological information, further support that the region anterior to *Hox1* in amphioxus is functionally wired to regulate Hox genes, whereas the region posterior to *Hox15* is not.


A
Amphioxus
1.5
1655
1739
1784
1801
2.0
2473
2.5
Mb
Meox
Copz1
Hnrnpab
Cbx
Smug1
Lnp
Gpatch8
Nfe2
Hnrnpaa
Skap
Mtx2
Evxa-Evxb
Hox cluster


B
Anterior
Skap
Mtx2
lncRNA
Hox1
Smug1
Repeats
Posterior
ScHox:
100 kb
2,400,000
2,450,000
2,500,000
Hox15
Slc20A-like
hyp. prot.
Evxa
Evxb
hyp. prot
Mdh1b
Lnp
Nifk
Msl3

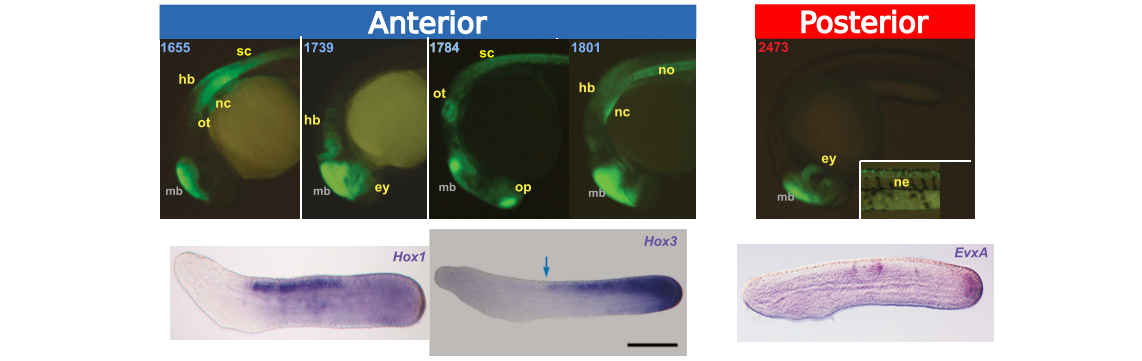

C
Anterior
Posterior
sc
hb
nc
ot
mb
ey
op
no
ne
Hox1
Hox3
EvxA

Figure 4.5: The anterior neighboring region, shared between amphioxus and vertebrates, was functionally wired to the regulation of Hox genes in the last common ancestor of chordates. (A) ATAC-seq in 36hpf embryos (track in black) allowed to identify four enhancers (in blue) in the anterior neighboring region included in the Hox TAD (marked with a blue rectangle) and one in the posterior region (red ATAC peak and rectangle). (B) Close up of the two regions. Again selected enhancers are colored and a track of repetitive elements is shown below. (C) Stable GFP transgenic embryos of 24hpf with the selected enhancers from both the anterior and the posterior regions. Below, RNA In-Situ Hybridization experiments in amphioxus embryos of the putative target genes of the different enhancers. On the left *Hox1* and *Hox3* in Pre-mouth amphioxus larvae showing collinear expression in the CNS (Pascual-Anaya et al. 2012). On the right *EvxA* showing expression in the neuroepithelium in a 36hpf embryo (Acemel et al. 2016). mb: midbrain (control of transgenesis), hb: hindbrain, sc: spinal cord, ot: otic vesicle, nc: notochord, ey: eye, oc: olfactory placode, ne: neuroepithelium

4.1.4

Hox promoters in the myriapod Strigamia maritima do not interact outside the cluster

We have shown that the anterior region flanking the only Hox cluster of amphioxus is syntenic with the anterior regions of the four mammalian clusters. In addition, this anterior region in amphioxus (but not the posterior) contains regulatory information utilized by Hox promoters. Therefore, a reasonable hypothesis is that amphioxus represents an intermediate state in the building of the 3D regulation of vertebrate Hox loci, with the anterior TAD linked to Hox genes, but missing connections with the posterior flanking region. In this scenario we are adventuring that this configuration was already present in the last chordate ancestor, from which both amphioxus and vertebrates evolved (Figure 4.6B). However, there are other plausible scenarios. For instance, the last chordate ancestor could have already had a bipartite configuration of the Hox locus, that was secondarily lost in amphioxus. In addition, the wiring of the anterior TAD could have happened before the last chordate ancestor: maybe at the root of deuterostomes or even earlier.

In order to further explore these alternatives we decided to look first at the synteny around the Hox clusters of non-chordate deuterostomes and also in protostomes. The synteny reconstruction of some representative organism are shown in Figure 4.6A. Interestingly, we found that homologs of some of the Hox neighboring genes in vertebrates could also be found in the vicinity of Hox genes in invertebrate genomes such as those of echinoderms, arthropods and mollusks. However, the ordering and the orientation was highly variable. In the sea star *Acanthaster plancii*, for example, the *Evx* gene is found in the anterior flanking region while *Copz* and *Nfe2* are present in the posterior one, in contrast to the situation in vertebrates. Similarly, in the limpet *Lottia gigantea Copz, Skap* and *Mtx2* are also found in the posterior flank. Another interesting example is the centipede *Strigamia maritima*, with *Lunapark* and *Hibadh* at the anterior end of the cluster. The lack of microsyntenic constraints between chordates and the rest of animals suggests that the Hox promoters of the last common ancestor of bilaterians did not rely much on regulatory inputs coming from either the anterior or the posterior flanking regions. Therefore, the wiring of the anterior flanking region is probably a chordate novelty.

With the intention of providing some experimental support to our syntenic analysis we decided to profile the 3D organization of the single Hox cluster of the myriapod *Strigamia maritima*. *Strigamia maritima* is an interesting model because, like amphioxus, displays a rather conservative genomic configuration of the Hox cluster compared to more common invertebrate models such as

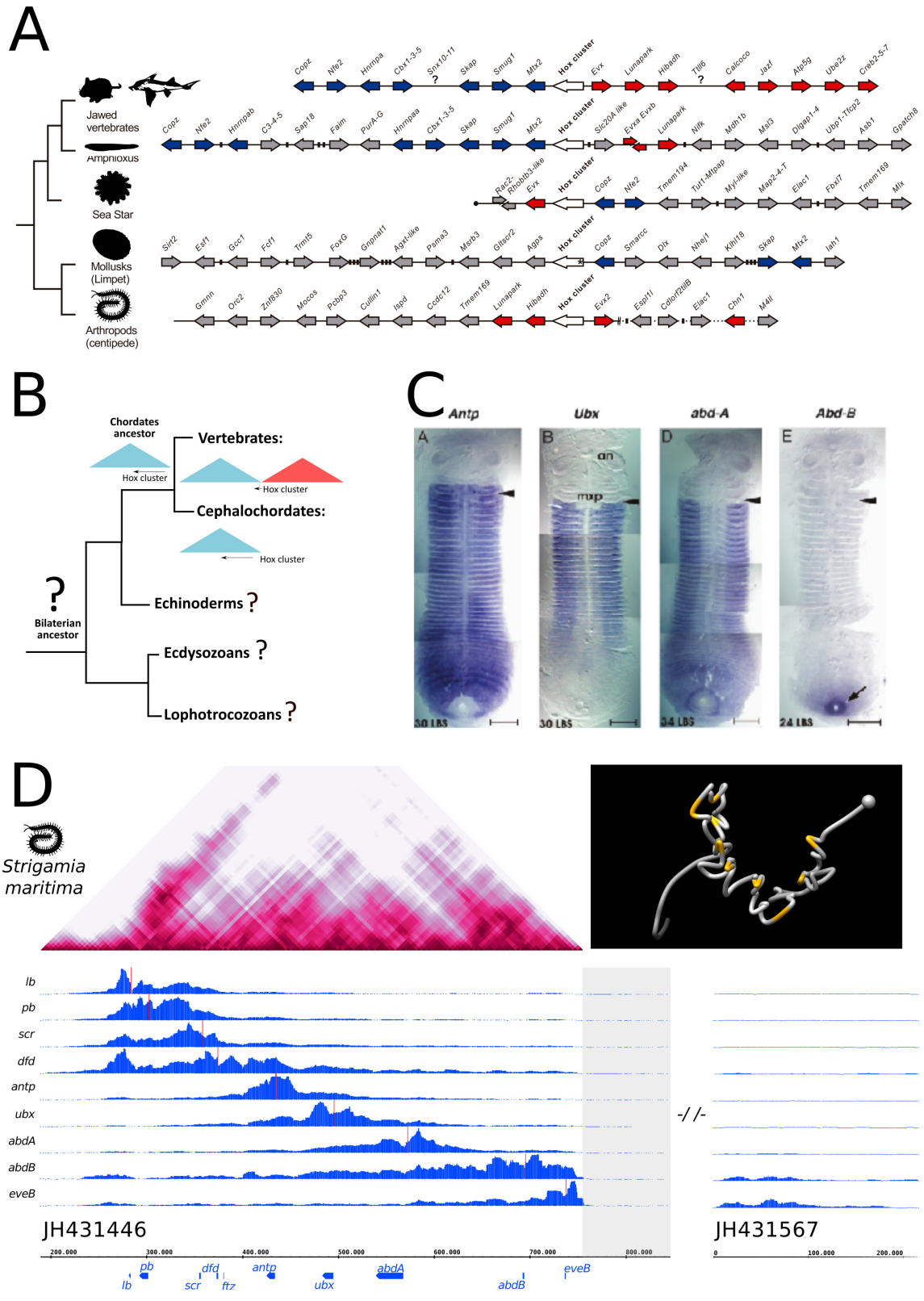

A
Jawed vertebrates
Amphioxus
Sea Star
Mollusks (Limpet)
Arthropods (centipede)
Hox cluster
B
Chordates ancestor
Vertebrates:
Hox cluster
Hox cluster
Cephalochordates:
Hox cluster
Echinoderms ?
Bilaterian ancestor
?
Ecdysozoans ?
Lophotrocozoans ?
C
Antp
Ubx
abd-A
Abd-B
A
B
D
E
an
mxp
30 LBS
30 LBS
34 LBS
24 LBS
D
Strigamia maritima
lb
pb
scr
dfd
antp
ubx
abdA
abdB
eveB
-//-
JH431446
JH431567
200.000
300.000
400.000
500.000
600.000
700.000
800.000
0
100.000
200.000
pb
dfd
antp
abdA
eveB
lb
scr
ftz
ubx
abdB

Figure 4.6: The neighboring region anterior to *labial* (*Hox1*) in *Strigamia maritima* does not interact with Hox promoters. (A) Syntenic reconstructions outside the chordate lineage do not reveal synteny conservation neither with the anterior nor the posterior vertebrate neighboring regions. (B) Phylogenetic tree showing the gaps in our knowledge of the chromatin architecture around Hox genes outside the chordate lineage. (C) *Strigamia maritima* displays a canonical Hox cluster, expressed collinearly in the segments of the embryo as shown by In-Situ Hybridization (Brena et al. 2006). *Antp* is expressed both in the maxilla (mxp) and in the leg bearing segments (LBS), *Ubx* is expressed only in LBS, *AbdA* is expressed both in LBS and in the proctadeum and finally *AbdB* is mostly expressed in the proctadeum. (D) 4C-seq derived model (left) and v-HiC (right) of the *Strigamia maritima* Hox locus. Below are the 4C-seq experiments used to derive the 3D-models.

Caenorhabiditis elegans or *Drosophila melanogaster*. The cluster occupies almost 400kb from *lab* to *abdB* and contains all the genes that were likely present in the last common ancestor of arthropods, with the exception of *Hox3*. Furthermore, *Strigamia* also shows the canonical collinear expression of the Hox genes along its embryonary segments (Figure 4.6C, Brena et al. 2006). Therefore we designed 4C-seq experiments from the promoters of eight Hox genes including *lb, pb, scr, dfd, ubx, antp, abdA* and *abdB* plus *eveB* (the *Evx* homolog). We performed these experiments in a pool of germ band stage embryos and the data were analyzed in an equivalent manner to those obtained from amphioxus and zebrafish (see Material and Methods, section 3.1.2, p.56). This included the 3D modelling of the locus and the generation of a virtual HiC heatmap (section 3.1.3, p.60). The first finding is that few if any contacts were established between anterior Hox genes (such as *lb, pb, scr* and *dfd*) and the anterior flanking region (Figure 4.6D). The frequency of interactions of these four promoters drop sharply towards the anterior end of the *lb* gene. This insulation seems to be slightly milder at the posterior end, with *abdB* being able to interact weakly with the posterior neighboring region. These regions can be found in a different scaffold of the *Strigamia* assembly, but the gap can be easily bridged thanks to the 4C-seq information. In contrast to the situation in amphioxus, these contacts involve regions that are not syntenic with vertebrates and it is likely that they were not established in the bilaterian ancestor.

Together with the lack of interactions at the anterior side (that is the one found wired in amphioxus) the most parsimonious model for the evolution of the vertebrate Hox 3D configuration is the following. First, no regulatory neighboring region contacted Hox promoters in the bilaterian ancestor. Secondly, in the last common ancestor of chordates, the anterior neighboring region was functionally incorporated. After that, and before the last common ancestor of vertebrates, a genomic rearrangement brought a different genomic region to the posterior end of the Hox cluster. Subsequently, this region was also incorporated and a TAD boundary was created. Finally, the Hox cluster was shrunk and reallocated precisely at the boundary. However, it is important to note that the mechanics, the timing and even the order of the events happening from the last chordate ancestor to the last vertebrate ancestor are still highly speculative.

4.2

Global 3D changes occurring during the evolution of the vertebrate body plan

4.2.1

Differences in size and regulatory content of developmental RLs along the evolution of chordates

In the previous section we elaborated that changes in the three dimensional arrangement of the chromatin around the Hox loci happened at the root of vertebrates. These changes were critical in order to establish a new strategy of transcriptional regulation for the vertebrate HoxA and HoxD genes in the patterning of limbs. We then wondered to what extent changes in the 3D architecture around other developmental genes were also important during the invertebrate to vertebrate transition.

In order to address this question we first decided to perform 4C-seq experiments around the promoters of a selection of orthologous developmental genes in two distant vertebrate species (24hpf zebrafish and E9.5 mouse embryos, the developmental stages are around the phylotypic stage) plus the cephalochordate amphioxus (15hpf embryos, early neurula). Since teleost fishes underwent an extra round of WGD, in principle this sampling allowed to look at the effect of RL duplication in chromatin architecture. The final set included 23 promoters in amphioxus, 63 in mouse and 107 in zebrafish, most of them either being transcription factors (e.g. FoxA, Dlx, Irx, Nkx and Meis) or upstream members of well known signaling pathways such as Hh, Fgf, Bmp or Wnt. The final composition of the set is showed in Table 4.1 and Table 4.2. Visual inspection of the resulting 4C-seq profiles of orthologous genes revealed noticeable changes in RL size between the three species. Perhaps not surprisingly, mouse promoters were able to interact with wider regions of chromatin than zebrafish and amphioxus. This is concordant with the increased genome size and intergenic distances in this species. Zebrafish RLs were also found to be bigger than those of amphioxus.

An illustrative example is the evolution of the RLs of the Dlx family that is shown in the Figure 4.7A. The promoter of the only *Dlx* gene in amphioxus interacted with a relatively small region of 200kb. In fact, a sharp decay in the interactions was visible downstream from the promoter (to the left in the figure), short before the locus of *En*. *En* is another important developmental gene presumably subjected to tight transcriptional regulation and this drop in the contacts may reflect the need to insulate *En* from *Dlx* regulation. In most vertebrates it is possible to find up to six genes orthologous to the amphioxus *Dlx* gene placed in three different loci (*Dlx1* to *Dlx6*). The latter is most probably due to a tandem duplication of the original *Dlx* gene followed by the two rounds of WGD and interestingly none of these loci is linked to any of the two vertebrate *En* paralogs. Furthermore, the *Dlx5-Dlx6* pair in mouse interacts with a long genomic region expanding almost one megabase downstream from *Dlx5*, including the genes *Shfm1* and *Slc5a13*.

Gene family	Amphioxus	Mouse	Zebrafish
BMP	-	Bmp2*	bmp2a* -
		Bmp4	bmp4*
		-	bmp7a bmp7b
DLX	Dlx*	Dlx1	dlx1a -
		Dlx2	dlx2a dlx2b
		Dlx3*	- dlx3b*
		Dlx4*	dlx4a* dlx4b*
		Dlx5*	dlx5a* -
		Dlx6*	dlx6a* -
EN	En	En1*	- -
		En2	eng2a eng2b*
FEZF	Fezf	-	fezf1* -
FGF	Fgf8	Fgf8	- fgf8b*
FOXA	FoxAa*	FoxA1*	foxa1
		FoxA2	foxa2*
FOXG	-	FoxG1	foxg1a*
			foxg1c foxg1d
FOXO	-	Foxo1	foxo1a foxo1b*
GBX	Gbx	Gbx2	gbx2*
GSC	Gsc	-	gsc*
GSX	Gsx	Gsx1*	gsx1
		Gsx2*	-
HAND	-	Hand2*	hand2
HEY	-	Hey2	hey2*
HH	Hh	Shh*	shha shhb
		Ihh*	-
		Dhh*	-
INSIG	-	Insig1	insig1*
		Insig2*	-
IRX	-	-	irx1a irx1b
			irx2a -
			irx3a* irx3b
			irx4a irx4b
			irx5a irx5b
			irx6a -
ISL	Isl*	Isl1*	isl1
		Isl2	isl2a* isl2b
MAF	-	Mafa	mafa
		Mafb*	mafba mafbb*
MEIS	-	Meis1*	- meis1b
MEOX	Meox	Meox1*	meox1
		-	- meox2b
MYC	-	Myc*	myca mycb*
MYF	-	-	myf5
			myf6
MYOG	-	Myog	myog*
NKX	Nkx2.1	Nkx2.1	nkx2.1
	Nkx2.2*	Nkx2.2	nkx2.2a* nkx2.2b
	-	Nkx2.4*	- nkx2.4b
	-	Nkx2.9*	-
	-	Nkx6.2	nkx6.2
NR2F	-	Nr2f2	-

Table 4.1: 4Cseq experiments by gene family (I). *Viewpoints with a single 4C-seq replicate.

Gene family	Amphioxus	Mouse	Zebrafish
OLIG	-	Olig1* Olig2*	olig1 olig2
OTP	Otp	-	-
OTX	Otx	Otx1 Otx2	otx1a otx1b* otx2
PAX2.5.8	Pax2.5.8	Pax2 Pax5 -	pax2a pax2b pax5* -
PAX3.7	Pax3.7	Pax3* Pax7	pax3a* - pax7a* pax7b*
PAX4.6	Pax4.6*	- Pax6*	- pax6a* pax6b*
PDX	Pdx	Pdx1	pdx1
PITX	-	Pitx2	pitx2
RAX	Rax	Rax	rx3*
SALL	-	Sall1 Sall4	sall1a - sall4*
SFRP	-	Sfrp1	sfrp1a sfrp1b
SIX	Six1-2* Six3-6* Six4-5*	- - - - -	six1a six1b six2a six2b six3a six3b six6a six6b six4a six4b
SOX	-	Sox11	sox11a* sox11b
TBX	-	Tbx2 Tbx3 Tbx4 Tbx5	- tbx2b tbx3a - tbx4 tbx5a* tbx5b*
TCF	-	Tcf7l2 -	tcf7l2 tcf21
WNT	-	Wnt1 Wnt8b* Wnt10b*	wnt1 wnt8a wnt8b wnt10a* wnt10b*
WT	-	-	wt1a wt1b

Table 4.2: 4Cseq experiments by gene family (II). *Viewpoints with a single 4C-seq replicate.

Importantly, the *Dlx6/Dlx5/Shfm1/Slc5a13* syntenic block could be traced in the zebrafish genome and is included within the RL of *dlx5a*. However, this genomic region is much smaller in zebrafish, occupying less than 500kb. RL size is an important feature since bigger RLs can potentially accommodate a larger number of cis regulatory elements, including enhancers. Therefore, we decided to find a way to calculate the extension of the RL of the genes automatically. We took advantage of the recently developed peakC R package (Geeven et al. 2018) that calls significant 4C-seq interactions. We used those significant interactions to develop a method that predicts the extension of RLs from the 4C-seq signal. We applied this method to a limited set of 4C-seq experiments, those where two replicates were available, since peakC relies heavily on replicates to calculate significance. This set included 12 promoters from amphioxus, 33 from mouse and 69 from zebrafish. We then decided to validate our method by comparing the RL prediction of mouse promoters with available HiC data from mESC (Figure 4.7B). Briefly, we accumulated the HiC signal surrounding our predicted RL boundaries and plotted the consensus HiC heatmap. The heatmap clearly showed two domains with a boundary at the center of the plot that coincided with a sharp

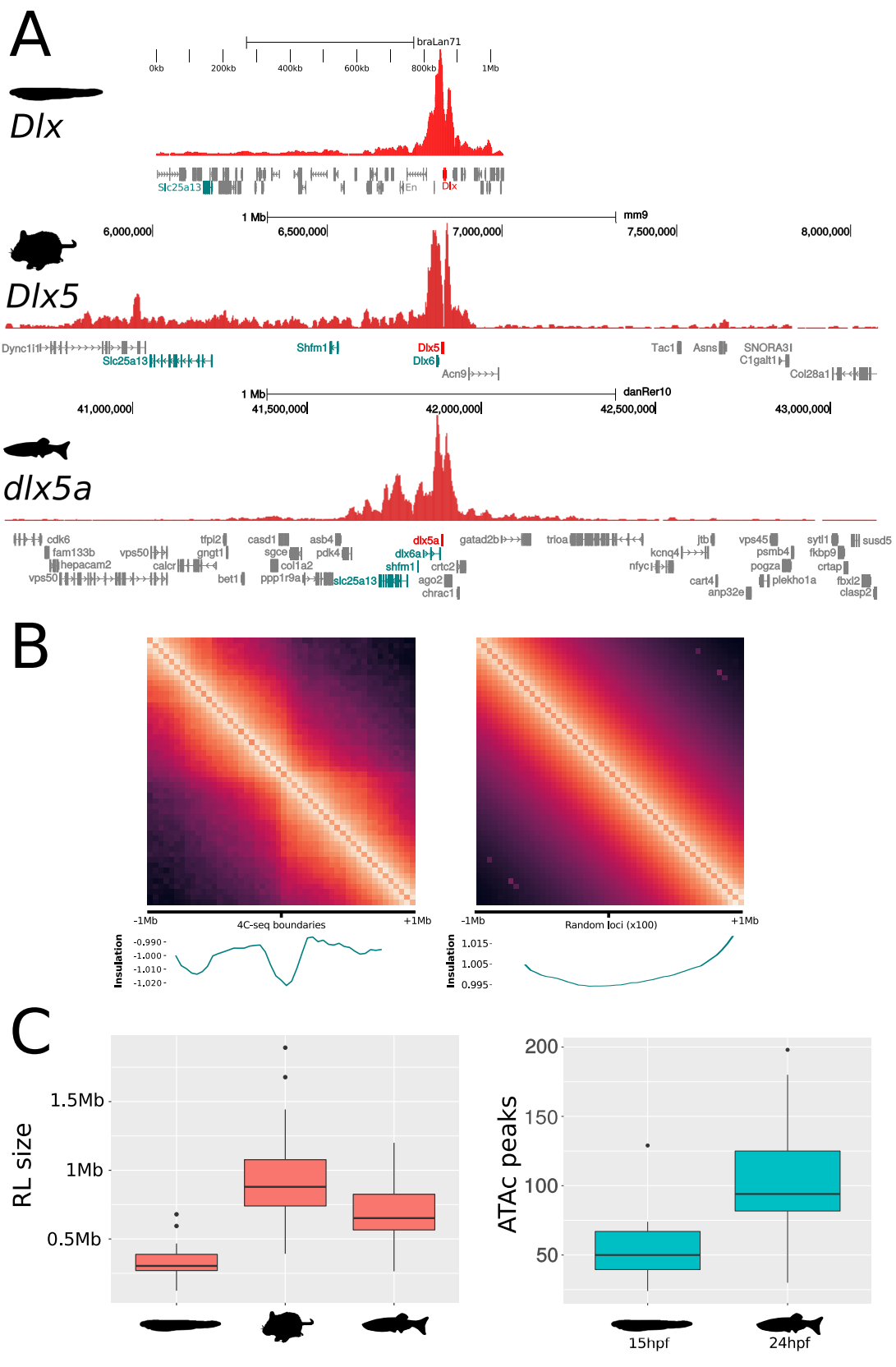

A
Dlx
braLan71
0kb
200kb
400kb
600kb
800kb
1Mb
Slc25a13
En
Dlx
Dlx5
1 Mb
mm9
6,000,000
6,500,000
7,000,000
7,500,000
8,000,000
Dync1i1
Slc25a13
Shfm1
Dlx5
Dlx6
Acn9
Tac1
Asns
SNORA3I
C1galt1
Col28a1
dlx5a
1 Mb
danRer10
41,000,000
41,500,000
42,000,000
42,500,000
43,000,000
cdk6
fam133b
hepacam2
vps50
tfpl2
gngt1
calcr
bet1
casd1
asb4
sgce
pdk4
col1a2
ppp1r9a
slc25a13
dlx5a
dlx6a
shfm1
crtc2
ago2
chrac1
gatad2b
trioa
jtb
kcnq4
nfyc
vps45
psmb4
pogza
cart4
anp32e
sytl1
susd5
fkbp9
crtap
plekho1a
fbxl2
clasp2
B
-1Mb
4C-seq boundaries
+1Mb
Insulation
-0.990
-1.000
-1.010
-1.020
-1Mb
Random loci (x100)
+1Mb
Insulation
1.015
1.005
0.995
C
RL size
1.5Mb
1Mb
0.5Mb
ATAc peaks
200
150
100
50
15hpf
24hpf

Figure 4.7: 4C-seq data reveal an increased size of the RLs of developmental genes in the vertebrate lineage. (A) 4C-seq experiments of different members of the Dlx family in amphioxus (*Dlx*, top), mouse (*Dlx5*, middle) and zebrafish (*dlx5a*, bottom). The names of the syntenic genes are in pale blue and the names of the 4C-seq baits in red. (B) Pile-up plots with the overall mouse ES cells HiC signal (Dixon et al. 2012) around 4C-seq derived RL boundaries (left) and around randomized RL positions (right, randomized 100 times). Insulation scores are calculated below. Further details in Material and Methods 3.1.2. (C) Global comparison of 4C-seq derived RL sizes between amphioxus, zebrafish and mouse (left) and enhancer content inside those RL (right, ATAC-seq peaks used as enhancer content proxy).

drop in the insulation score that correlates inversely with the boundary strength. We compared these results with those obtained when randomizing the location of the RLs. No insulation could be detected in the consensus plot of randomized boundaries, neither a sharp drop of the insulation score. Additional details about the method and the validation, including the code, is available in the corresponding Material and Methods section (3.1.2, p.56).

When we compared the predicted sizes in our set we readily found significant differences between the three species (Figure 4.7C left, Kruskal-Wallis p-value = 1.423 x 10^{-8}). Pair-wise comparisons revealed that amphioxus RLs were significantly smaller than those of zebrafish and mouse (Wilcoxon tests p-values of 6.571 x 10^{-6} and 9.198 x 10^{-9} respectively). In addition, zebrafish RLs were also significantly smaller than those in mouse (p-value = 1.394 x 10^{-4}). Lastly, we decided to check if indeed bigger RLs in vertebrates with respect to amphioxus were indicative of the presence of more regulatory information in vertebrate RLs. In order to do that we took advantage of ATAC-seq experiments done both in 15hpf embryos of amphioxus and 24hpf embryos of zebrafish (equivalent stages to those used for the 4C-seq experiments, **Marletaz2019**). We quantified the number of open chromatin regions (predictive of the presence of cis-regulatory elements and enhancers) located inside our predicted RLs and we found that zebrafish RLs on average doubled the amount of regulatory information that could be found in amphioxus RLs (Wilcoxon p-value = 1.821 x 10^{-5}, Figure 4.7C right). These results suggest an increase in regulatory complexity of RLs in the vertebrate lineage, at least in the selected set of developmental genes.

4.2.2

Current RLs are composed of genomic strata wired in different reorganization events

One of the most surprising observations from the *Dlx5* 4C-seq experiments was that 3D interactions both in zebrafish and mouse decayed at equivalent syntenic locations (soon after the gene *Slc5a13*). This is evident even though the sizes of those syntenic regions are strikingly different between the two species. Perhaps an even clearer example arise from the comparison of the mice *Insig1/En2/Shh* locus with the zebrafish *insig1/eng2a/shha* locus. In mouse, these three genes are split in two TADs harboring the *Insig1-En2* pair and *Shh* respectively, as shown by mESC HiC (Figure 4.8A). The *Insig1/En2* TAD starts near the gene *Htr5a* and ends in the vincinity of *Shh*. Besides, the big *Shh* TAD ends more than a megabase away from this gene, near the *Ube3c* and *Mnx1* loci. The 3D configuration in two TADs in mouse is readily visible using the 4C-seq profiles of the promoters of *Insig1*, *En2* and *Shh* in E9.5 embryos, with the frequencies of interaction dropping sharply beyond the boundaries of the mESC TADs (Figure 4.8A). Importantly, frequent contacts were observed between *Shh* and the introns of the gene *Lmbr1*, that contains the well known limb enhancer

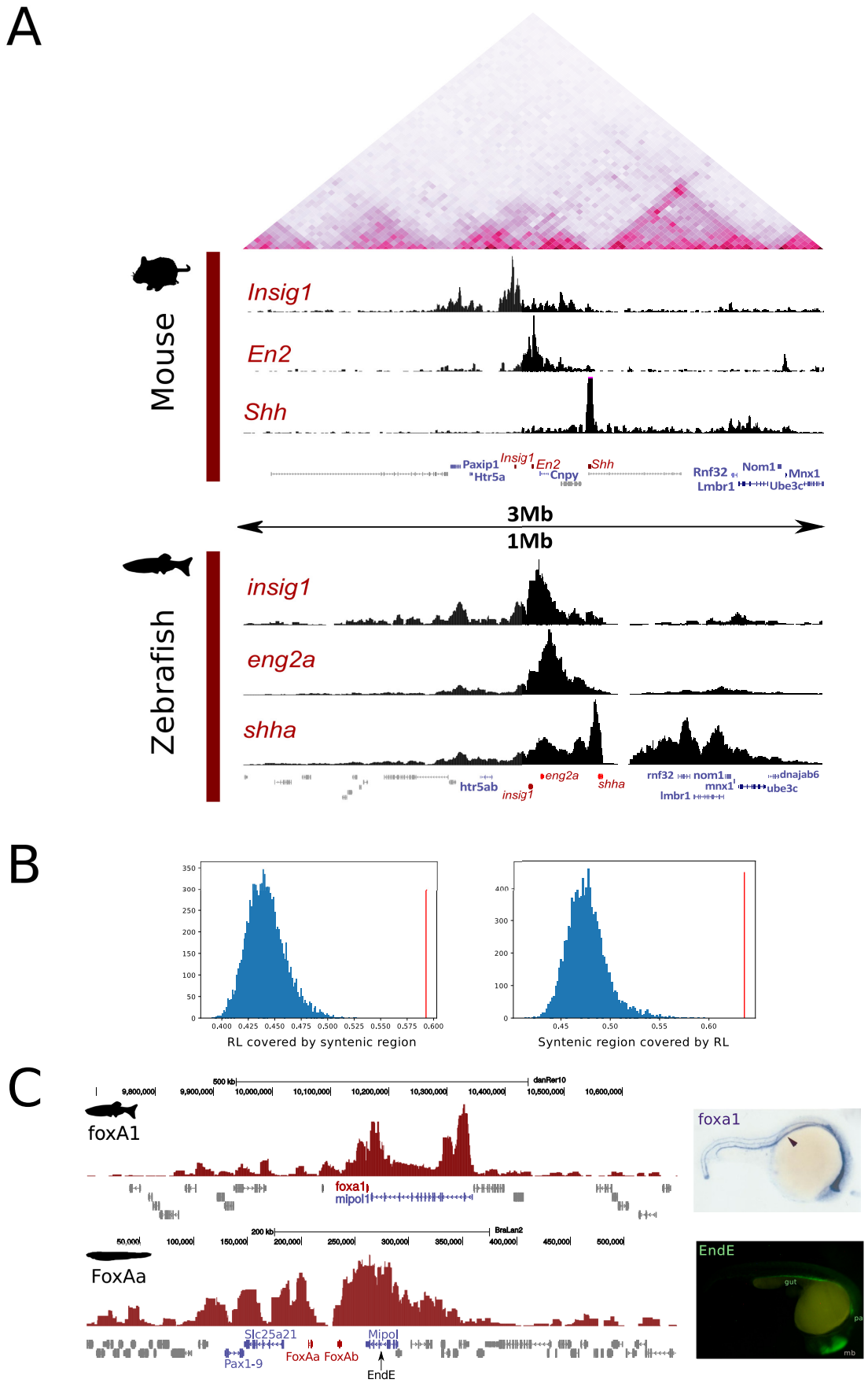

A
Mouse
Insig1
En2
Shh
Paxip1
Htr5a
Insig1
En2
Cnpy
Shh
Rnf32
Nom1
Mnx1
Lmbr1
Ube3c
3Mb
1Mb
Zebrafish
insig1
eng2a
shha
htr5ab
insig1
eng2a
shha
rnf32
nom1
dnajab6
mnx1
lmbr1
ube3c
B
RL covered by syntenic region
Syntenic region covered by RL
C
foxA1
foxa1
mipol1
foxa1
FoxAa
Slc25a21
Pax1-9
FoxAa
FoxAb
Mipol
EndE

Figure 4.8: RLs tend to overlap with syntenic blocks conserved from teleosts to mammals. (A) Mouse ES cell HiC around the *Shh* locus (Dixon et al. 2012, top) and 4C-seq experiments in mouse E9.5 embryos (middle) and in the zebrafish orthologous loci (24hpf embryos). The names of the genes used as 4C-seq baits are shown in red, and the names of additional genes conserved in synteny between zebrafish and mouse in blue. (B) Plots with the distribution of the expected overlap between RLs and automatically predicted syntenic blocks, in both directions (see Material and Methods, section 3.1.2). The red lines represent the experimentally derived overlap. (C) The ancient RL stratum of *FoxA*, that is shared between the amphioxus *FoxAa* and the zebrafish paralog *foxA1*. 4C-seq contacts populate the introns of the neighboring syntenic gene *Mipol* in both cases. The introns of the amphioxus *Mipol* gene contain a functionally conserved endodermal enhancer (EndE) that drives expression in zebrafish to endodermal territories compatible with the expression pattern in zebrafish (mb: midbrain; pa: pharyngeal arches). The In-Situ Hybridization experiment in zebrafish was published elsewhere (Alexander et al. 1999). The transgenesis was performed as explained in Material and Methods section 3.4

ZRS. A very similar scenario can be found when exploring the 4C-seqs of the equivalent locus in 24hpf zebrafish embryos. First, the interactions of *insig1* and *eng2a* are highly overlapping and comprehended between the gene *htr5ab* and *shha* (Figure 4.8A). In addition, *shha* interactions extend up to the *mnx1-ube3c* pair of genes, equivalently to the situation in mouse. Furthermore, the general interaction profiles of the three orthologous genes are qualitatively very similar between zebrafish and mouse. However, the distance between *Htr5a* and *Shh* that roughly delimits the *Insig1/En2* TAD is approximately 600kb. In contrast, the distance between shha and *htr5ab* in zebrafish is 200kb, three times shorter. Similarly, more than 1 Mb separates *Shh* from the *Ube3c* gene in mouse while the distance between *shha* and *ube3c* is 250kb. This might reflect that TADs are very plastic to either the gain or the loss of DNA sequences, while being specially sensitive to mutations occurring near the boundaries.

In order to further explore the relationship between synteny and RLs we combined our automatic prediction of RLs from 4C-seqs and looked for the overlap between those RLs with syntenic blocks calculated automatically from available chain alignments between the zebrafish and the mouse genomes. Interestingly, we found that 59% of the extension of our zebrafish RLs was syntenic with mouse. At the same time, we also found that 64% of the syntenic regions involving our baits were covered by the bait RL. Both proportions proved to be above what could be expected by chance when using random perturbations of the RLs (Figure 4.8B, empiric p-value $< 1 \times 10^{-5}$, details about the RL randomization in Material and Methods, 3.2: p.58). Such analysis, however, was not feasible including the amphioxus RLs due to the relatively poor contiguity of the assembly and the limited set of amphioxus 4C-seqs available.

However, it was possible to manually identify equivalent syntenic regions inside the RL of orthologous genes when comparing vertebrates and amphioxus. Such syntenic regions have presumably contributed to the regulation of the same genes from the last common ancestor of chordates and one of them is related to the *FoxA* gene. *FoxA* and his neighbor, *Mipol1*, are found together both in vertebrates and in amphioxus. In addition, when inspecting the 4C-seq profiles of *FoxAa* in amphioxus and *foxa1* in zebrafish it was possible to establish that *Mipol* and *mipol1* respectively are included within the RL of their respective nieghboring FoxA genes (Figure 4.8C). Furthermore, we decided to test the enhancer potential of an open chromatin region detected inside one of the *Mipol* introns of amphioxus (termed EndE in the Figure 4.8C) using transgenic reporter assays in zebrafish. Strikingly, this enhancer drove expression in several endodermal territories in zebrafish that were compatible with the endogenous pattern of expression of *foxA1*, including the gut and the pharyngeal arches. This suggests that this ancient syntenic region also contains ancient and

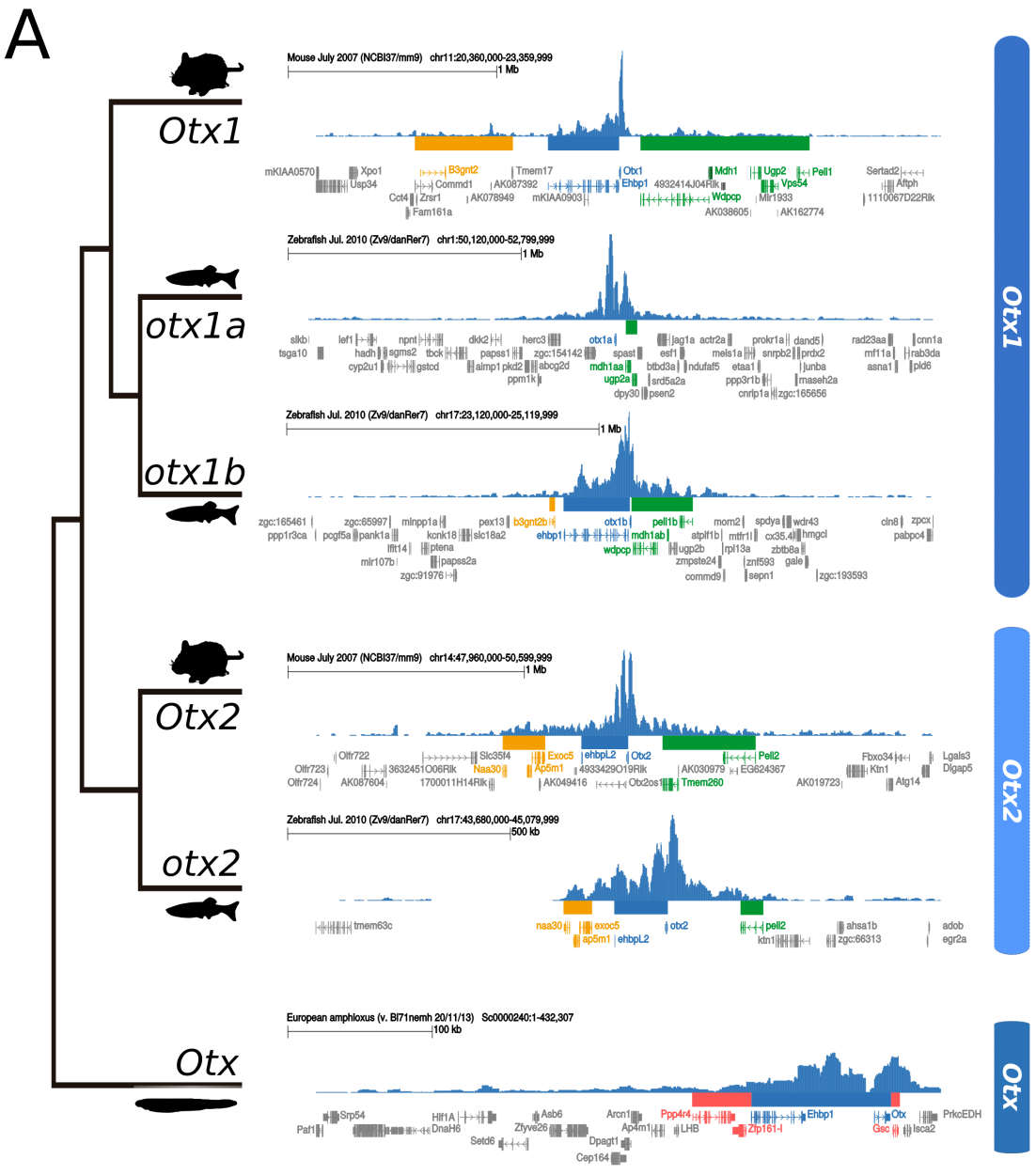

A
Otx1
otx1a
otx1b
Otx2
otx2
Otx
Otx1
Otx2
Otx
Mouse July 2007 (NCBI37/mm9) chr11:20,360,000-23,359,999
Zebrafish Jul. 2010 (Zv9/danRer7) chr1:50,120,000-52,799,999
Zebrafish Jul. 2010 (Zv9/danRer7) chr17:23,120,000-25,119,999
Mouse July 2007 (NCBI37/mm9) chr14:47,960,000-50,599,999
Zebrafish Jul. 2010 (Zv9/danRer7) chr17:43,680,000-45,079,999
European amphioxus (v. BI71nemh 20/11/13) Sc0000240:1-432,307
B
B. lanceolatum
S. kowalewskii
L. anatina
Gpb5
Gpa2
Ppp4r4
Zfp161-like
Ehbp1
Otx
Gsc
Isca2
PrkcEDH
Ig-like
Pou2*
Boc-Cdon*
Figla-like
Eif3k
2700097O09Rik

Figure 4.9: The actual RLs of the different members of the *Otx* family are composed of strata of different ages. (A) Complete set of 4C-seqs of the paralog RLs of the *Otx* family in amphioxus (*Otx*, bottom), zebrafish (*otx1a, otx1b* and *otx2*) and mouse (*Otx1* and *Otx2*). The blue rectangles mark the ancient *Ehbp/Otx* syntenic block shared by all members of the family. The green rectangles mark the *Otx/Peli* syntenic block present in all vertebrate copies. Yellow rectangles mark paralog specific syntenic regions shared by either the *Otx1* or the *Otx2* vertebrate paralogs. Finally, the red rectangle in the amphioxus track shows an ancient syntenic block including the *Ppp4r* gene and *Gsc* that is conserved to the LCA of bilaterians but was lost secondarily in vertebrates. A simplified syntenic reconstruction of this ancient block is shown in (B).

functionally conserved regulatory sequences.

Importantly, despite the constraints to keep the integrity of RLs, a considerable proportion of them (around the 40% when comparing zebrafish with mouse) are composed of newly wired regions. This proportion is much higher if we compare those two vertebrate species with amphioxus. An illustrative example of the wiring of new genomic regions is the evolution of the RLs of the Otx family of TFs (Figure 4.9A). In amphioxus there is only one copy of this TF, while in most vertebrates there are two paralogs that were retained from the hypothetical four copies that originated from the two rounds of WGDs (*Otx1* and *Otx2*). In zebrafish there are three paralogs due to the teleost specific round of WGD, one of them related to *Otx2* (*otx2*) and the other two related to *Otx1* (*otx1a* and *otx1b*). Similarly to the case of *FoxA* and *Mipol*, an ancient portion of the RL can be traced back to the root of bilaterians thanks to the syntenic gene *Ehbp* (the region is indicated with blue rectangles in the Figure 4.9A). This ancient region constitute a big proportion of the RLs of the amphioxus *Otx*, of the *Otx1* and *Otx2* paralogs in mice and of the *otx2* and *otx1b* paralogs in zebrafish. However, this region could not be distinguished in the *otx1a* zebrafish paralog. Either the region was replaced or the exons of the *Ehbp* gene have been completely erased. An example of the second scenario was the case of the neighboring genes of the vertebrate Hox clusters shown earlier. Nevertheless, the 4C-seq profiles and the conservation of synteny indicated that the *Otx/Ehbp* region was wired to the *Otx* promoter at the very least in the common ancestor of chordates. If we use bare synteny as a proxy, such configuration appeared even earlier in the last common ancestor of bilaterians. However, 4C-seq or HiC data from non chordate species would be needed to confirm this (Figure 4.9B). Apart from this highly conserved stratum, it was also possible to identify a younger syntenic region that was included inside the RL of all the different vertebrate paralogs of *Otx*. However, this region is absent in amphioxus. It can be identified thanks to the *Wdpcp/Tmem260/Mdh1/Ugp2/Vps54/Peli* syntenic block (depicted with green rectangles in the Figure 4.9A). Since this stratum is integrated inside the RL of *Otx* both in the *Otx1* and the *Otx2* lineages, the most parsimonious explanation is that it got wired before the two rounds of WGDs, in the common ancestor of vertebrates. In addition, we found another two regions that are wired specifically either to the *Otx1* or to the *Otx2* paralog lineages (those are depicted with yellow rectangles). In the case of the *Otx1* lineage the syntenic block is *Otx1/Ehbp1/B3gnt2* and in the case of the *Otx2* lineage the block is *Otx2/EhbpL2/Exoc5/Ap5m1/Naa30*. Therefore, these strata must have become connected to the *Otx* regulation after the two rounds of WGDs. Finally, it is important to note that the ancestral syntenic block in the bilaterian ancestor was likely *Ppp4r4/Zfp161-l/Ehbp/Otx/Gsc*, and this complete block can be found in amphioxus connected to the *Otx* promoter (this extended bilaterian syntenic block is depicted with red rectangles in Figure 4.9A). The syntenic relationships between *Otx* and both the *Ppp4r4/Zfp161* pair of genes and *Gsc* must have disappeared at the root of vertebrates and were substituted by both the green and the

yellow genomic regions. In summary, the RLs of the *Otx* paralogs in vertebrates are composed by genomic regions that got wired in different moments of the evolutionary history of vertebrates (i.e. the *Ehbp/Otx* regions was present at the root of bilaterian animals, and subsequently several rearrangement episodes configured the final architecture of the locus of the vertebrate lineage). Each of those rearrangements could have potentially triggered the evolution of the pattern of expression of *Otx* by bringing new enhancers to the vincinity of this developmental gene. In following sections we will try to explore further this potential evolutionary mechanism using genome wide approaches.

4.2.3

HiChIPs against H3K4me3 allow to identify RLs genome wide both in zebrafish and in amphioxus

By using our 4C-seq experiments we were able to start inferring patterns of how RLs have evolved during the vertebrate origin and radiation. We were able to detect constraints in the location of RL boundaries within vertebrates and the flexibility of these RLs to changes in size. In addition, we found that despite the fact that RL boundaries are constrained, new genomic regions can be incorporated over time. However, the approach was limited by the selection of a specific subset of orthologous promoters, since performing 4C-seqs for all the promoters in a genome is not a reasonable strategy. Therefore we decided to explore the evolution of the 3D architecture genome wide using HiChIP.

In order to look at RLs we chose to perform HiChIP experiments against the H3K4me3 epigenetic modification that is deposited in all active promoters. We generated H3K4me3 HiChIP libraries both in 24hpf zebrafish and 15hpf amphioxus embryos to reveal the RLs of all active genes at once with high resolution. First, we visually inspected the contact matrices generated in zebrafish in comparison with publicly available HiC data also from 24hpf zebrafish embryos at 5kb resolution (Figure 4.10A, top). In the HiC data, dark triangles representing TADs started to appear although despite the sparsity of the data at the required resolution. In contrast, finely resolved structural information could be found around active promoters in the HiChIP heatmaps, with discrete lines or stripes emerging from active genes and delineating the different RLs. Importantly, a conventional H3K4me3 ChIP-seq signal can be extracted from the HiChIP experiments if we do not consider the pairing of the reads. This signal was consistent with the signal of conventional H3K4me3 ChIP-seq experiments performed in zebrafish embryos from the same stage (more details in Material and Methods section 3.2.2). As expected, the different HiChIP stripes emerged from H3K4me3 rich areas located around active promoters (Figure 4.10A).

Next, we decided to contrast the 3D information derived from the HiChIP experiments with our available zebrafish 4C-seq experiments. High qualitatively agreement was observed when comparing individual loci, for instance in the case of the RL of the zebrafish *myca* gene (Figure 4.10A). This agreement also holds when comparing the HiChIP with the entire collection of zebrafish 4C-seq experiments at 10kb resolution (Figure 4.10B). Both Pearson and Spearmann correlations between the different type of experiments were high (above 0.7 and 0.6 respectively, Figure 4.10C).

Finally, in order to be able to compare the HiChIP experiments from both species we developed an automatic algorithm to predict the extension of the RLs based on the HiChIP signal. The algorithm, that we termed *Sushibox*, scans all the positions in the genome and assigns them to the

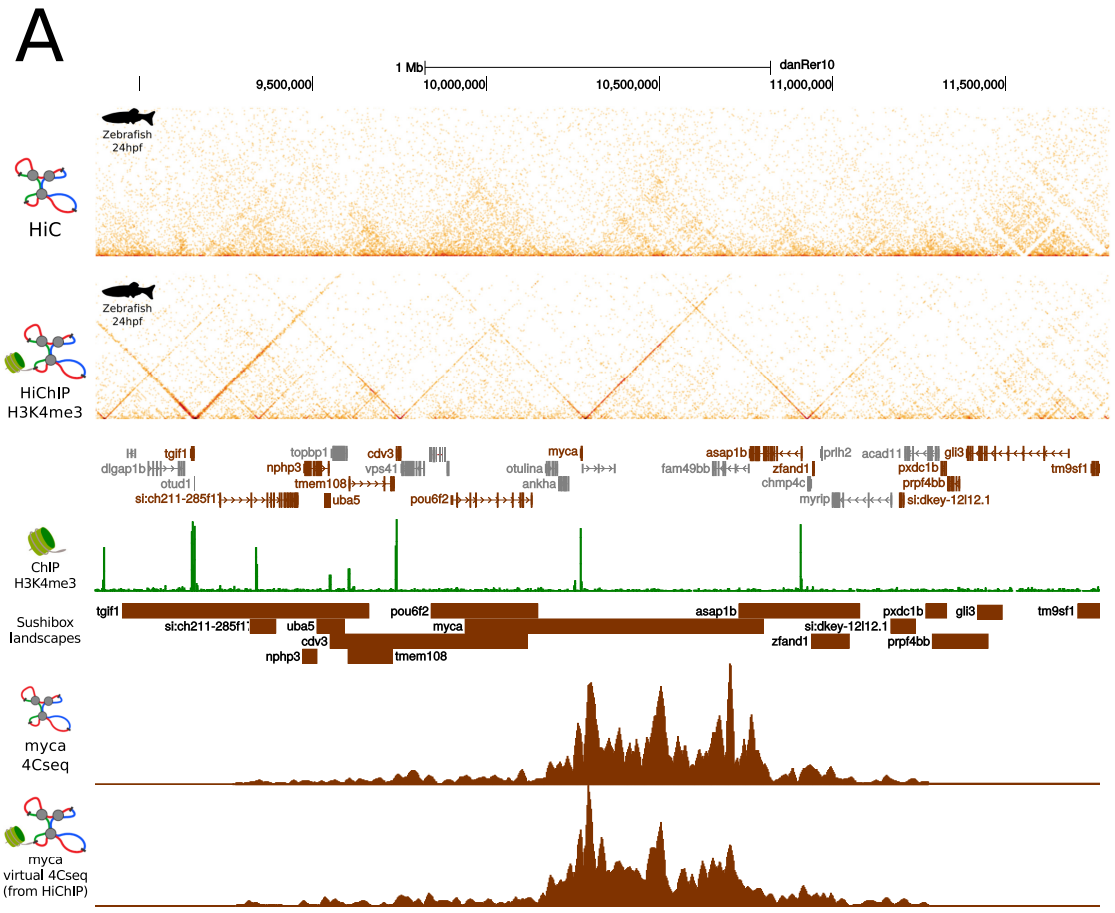

A
1 Mb
danRer10
9,500,000
10,000,000
10,500,000
11,000,000
11,500,000
Zebrafish 24hpf
HiC
Zebrafish 24hpf
HiChIP H3K4me3
tgif1
dlgap1b
otud1
si:ch211-285f11
topbp1
nphp3
tmem108
cdv3
vps41
uba5
pou6f2
myca
otulina
ankha
asap1b
fam49bb
zfand1
chmp4c
prlh2
myrip
acad11
gli3
pxdc1b
prpf4bb
si:dkey-12l12.1
tm9sf1
ChIP H3K4me3
Sushibox landscapes
myca 4Cseq
myca virtual 4Cseq (from HiChIP)
B
4Cseq rep1
4Cseq rep2
HiChIP rep1
HiChIP rep2
C
Pearson
Spearmann
HiChIP_rep1
HiChIP_rep2
4Cseq_rep1
4Cseq_rep2

Figure 4.10: Developmental RLs could be inferred using H3K4me3 HiChIP experiments. (A) Comparison between publicly available HiC data from 24hpf zebrafish embryos (top, Kaaij et al. 2018) and our H3K4me3 HiChIP data at the equivalent stage (below). The signal in the HiChIP is concentrated around active promoters. The ChIP-seq track below is derived from the same HiChIP experiment as described in the Material and Methods section 3.2.3. RL predictions are derived from the HiChIP contacts using the *Sushibox* algorithm (depicted with brown boxes below the ChIP-seq track, see Material and Methods section 3.2.2). Two bottom tracks show a comparison between a 4C-seq experiment using the gene *myca* as the bait and a v-4C track obtained directly from the HiChIP contact matrix. (B) Heatmap with qualitative comparisons between available 4C-seq experiments performed in zebrafish and the contacts extracted from the H3K4me3 HiChIP in equivalent loci at 10kb resolution. (C) Pearson and Spearmann correlation matrices obtained from the heatmaps in (B).

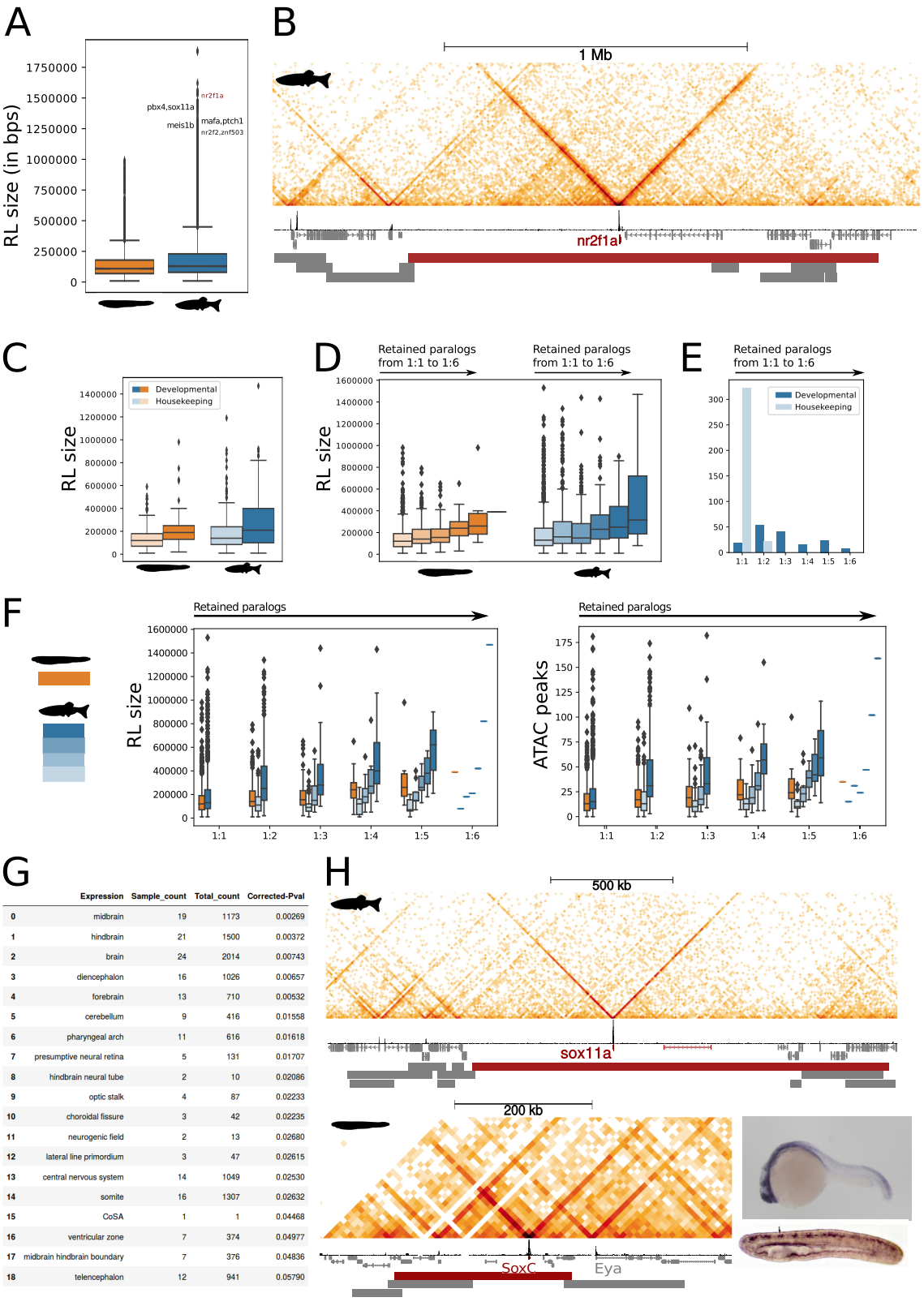
more likely active genes. A detailed description of the algorithm is provided in the Materials and Methods section 3.2.2, and an example of the resulting RL calling is depicted with brown boxes in the Figure 4.10A. Using this method we were able to identify 5612 RLs in amphioxus and 10133 RLs in zebrafish, a much more comprehensive collection than the previously derived using 4C-seq experiments.

4.2.4

RL evolution driven by WGDs and genomic rearrangements

Once we were able to automatically establish the location of the different RLs both in zebrafish and in amphioxus in our H3K4me3 HiChIP experiments, the way was paved to start exploring the evolution of those RLs at the onset of vertebrates. First of all we decided to explore the role of WGDs in the evolution of the vertebrate duplicated RLs, and we started to look at how the sizes of those RLs changed. Surprisingly, we found that the median size of the RLs does not differ much between zebrafish and amphioxus despite the fact that the genome size of zebrafish is twice as big (Figure 4.11A). These median sizes are relatively small in both species, which reflects that an important proportion of the promoters are not very responsive to the regulation by distal enhancers. However, it is worth noting that the upper tail of the distribution in zebrafish contain more RLs and that those RLs reach bigger sizes, above 1Mb in many cases. Furthermore, the biggest RLs in zebrafish are enriched in well known developmental regulators such as *sox11a, pbx4, meis1b* or *nr2f1a* (Figure 4.11B). Indeed, this pattern can be also observed in the opposite direction, with developmental regulators consistently displaying bigger RLs both in zebrafish and amphioxus (Figure 4.11C, developmental genes were defined as done in Marlétaz et al. 2018, Mann-Whitney U p-values of 3.2 x 10^{-5} and 1.8 x 10^{-7} for amphioxus and zebrafish comparisons respectively).

After this general descriptive observations we decided to explore the effect of WGDs by stratifying the analysis on the RL sizes by the number of paralog genes retained in zebrafish. For that purpose we first wanted to limit our analysis to families of ohnolog RLs, which are the paralogs that originated from the WGD events. As a proxy, we considered ohnologs those genes that are retained in 1 to 1, 1 to 2 , 1 to 3 or 1 to 4 proportions when comparing the amphioxus and the mouse genomes. Then we filtered only those ohnolog families where all members displayed H3K4me3 peaks at their promoters so all RLs could be reliably calculated by *Sushibox*. Applying such filters we ended up with a collection of 1678 zebrafish and 1195 amphioxus RLs that we used for further analysis. We observed a clear tendency: those genes that are retained in a higher number of copies in zebrafish tend to display bigger RLs both in zebrafish and in amphioxus (Figure 4.11D, Spear-



	Expression	Sample_count	Total_count	Corrected-Pval
0	midbrain	19	1173	0.00269
1	hindbrain	21	1500	0.00372
2	brain	24	2014	0.00743
3	diencephalon	16	1026	0.00657
4	forebrain	13	710	0.00532
5	cerebellum	9	416	0.01558
6	pharyngeal arch	11	616	0.01618
7	presumptive neural retina	5	131	0.01707
8	hindbrain neural tube	2	10	0.02086
9	optic stalk	4	87	0.02233
10	choroidal fissure	3	42	0.02235
11	neurogenic field	2	13	0.02680
12	lateral line primordium	3	47	0.02615
13	central nervous system	14	1049	0.02530
14	somite	16	1307	0.02632
15	CoSA	1	1	0.04468
16	ventricular zone	7	374	0.04977
17	midbrain hindbrain boundary	7	376	0.04836
18	telencephalon	12	941	0.05790

Figure 4.11: WGDs impacted the evolution of vertebrate RLs. (A) Boxplot with the comparison of the RL sizes predicted by *Sushibox* between zebrafish and amphioxus. (B) Contact matrix around the RL of the *nr2f1a* promoter in zebrafish, one of the biggest. (C) Developmental genes tend to have bigger RLs both in zebrafish and in amphioxus. (D) Ohnolog families that were retained in higher number of copies in vertebrates tend to have bigger RLs both in zebrafish and in amphioxus. (E) Ohnolog families retained in high number of copies are also enriched in developmental genes. (F) RL sizes stratified by the number of ohnologs retained and ordered by size reveal that when ohnologs are retained in more than one copy one of them tend to retain the original size and others grow both in size and in enhancer content. (G) Table with the expression domains enriched in those RLs that grew in the vertebrate lineage. Neural tissues and vertebrate novelties such as the pharyngeal arches are enriched. (H) The gene *sox11a* in zebrafish is one of the examples of expanded RL and the expression domains are neural both in zebrafish (Howe et al. 2012) and in amphioxus (Lin et al. 2009) embryos of 24hpf and 36hpf respectively.

mann's ρ p-values of 6.0 x 10^{-10} and 3.4 x 10^{-11} respectively). This pattern seems to be explained by the fact that genes that were retained in more than one copy after the WGD events are mostly developmental regulators (Figure 4.11E). Then, we wondered if after the WGDs the sizes of the different ohnolog RLs evolved indistinctly or if there was some kind of constraint or bias. In order to explore this we split and ranked the different zebrafish ohnologs by size and compared the sizes with the size of the RL of the amphioxus copy. By doing this we found that in the cases where only one ohnolog has been retained the RL size tend to be slightly bigger in zebrafish. In contrast, when more than one copy is retained, at least one of the copies tend to display a size that is as small or even smaller than the amphioxus copy. Meanwhile, the other zebrafish copies tend to exhibit RL sizes that are clearly bigger than the one from amphioxus (Figure 4.11F). Importantly, if we use the RLs to assign ATAC peaks (that are proxies for enhancers) to ohnologs and we quantify them we obtain an equivalent pattern. Although in principle several scenarios could be plausible, we assume that most of the times the size of the RL size in amphioxus would recapitulate the ancestral state in the LCA of chordates. Therefore, we interpret that after the WGDs it was common that one of the vertebrate ohnologs retained the ancestral size while the others were free to expand and perhaps increment their regulatory complexity.

We then decided to explore what were the expression patterns of those genes that expanded their RLs in the vertebrate lineage by taking advantage of the ZEOGs tool (Prykhozhij, Marsico, and Meijsing 2013, Material and Methods, section 3.2.2). Excitingly, we found a strong enrichment in expression terms related with neural development and also with the development of the pharyngeal arches which are a vertebrate novelty (Figure 4.11G). Then we sought to manually inspect some of the ohnologs that showed this kind of pattern like the SoxC family of TFs that comprises the *SoxC* gene in amphioxus and the *sox4a/sox4b* and *sox11a/sox11b* genes in zebrafish. The RL of the *SoxC* gene in amphioxus is 250kb, which is well above the median for that organism. In zebrafish, the sizes of the RLs range from the 150kb of *sox4b* (smaller than the amphioxus landscape) to the 1.4 Mb of *sox11a* (Figure 4.11H). Both the amphioxus *SoxC* gene and the orthologs in zebrafish are expressed in neural territories.

Next we wanted to investigate to what extent genomic rearrangements encompassing TAD boundaries like inversions, deletions or translocations could have caused GRN rewirings at the root of vertebrates. Given the big phylogenetic distance between zebrafish and amphioxus we focused our synteny analysis on syntenic pairs and included a number of additional organisms inspired by previous studies (Irimia et al. 2012). Apart from zebrafish and the european amphioxus species, we chose four additional vertebrates (medaka, chicken, mouse and human) and three non-vertebrate deuterostomes (the Asian amphioxus species *Branchiostoma belcheri*, the sea

urchin *Strongylocentrotus purpuratus* and the hemichordate *Saccoglossus kowalewskii*). In short, by exploring the different assemblies, we defined a linkage age category for every pair of neighboring genes both in zebrafish and amphioxus. In other words, we tried to infer for how long each pair of zebrafish and amphioxus genes have remained together. In the case of zebrafish we defined four categories: zebrafish specific, vertebrate specific, chordate specific or deuterostome root. In the case of amphioxus we had amphioxus specific, chordate specific and deuterostome root. It is important to note that the temporal resolution is limited by the number of available species. Then, syntenic pairs marked as zebrafish specific could have appeared anytime after the split between the zebrafish and the medaka lineages. Equivalently, syntenic pairs marked as deuterostome root could potentially be more ancient because we have not explored any protostome species. In order to achieve this classification we used publicly available gene annotations and the orthology prediction extracted from OMA as performed in Marlétaz et al. 2018 (further details, files and code are available in the Material and Methods section 3.2.2).

Once we defined the categories for every syntenic pair we started to compare this information with the RLs calculated from the H3K4me3 HiChIP experiments using *Sushibox*. TAD boundaries and therefore RLs has been described elsewhere to be evolutionarily stable, with breakpoints for genomic rearrangement often being placed precisely in TAD boundaries (Dixon et al. 2012, Vietri Rudan et al. 2015). We wanted to test if these constraints hold for long evolutionary distances. We hypothesized that if true, the longer two genes have remained together, the most likely is that they are encompassed by the same RL. This is indeed the pattern observed both using zebrafish and amphioxus RLs (Figure 4.12A). Syntenic pairs present from the deuterostome root are found together in the same RL 70.6% and 92.1% of the times in zebrafish and amphioxus respectively. These percentages drop to 68.0% and 85.0% respectively for chordate specific pairs and down to 38.1% and 59.% for the newest pairs. Lastly, vertebrate specific pairs share RL 59.0% of the times, which is an intermediate proportion in between chordate and zebrafish specific pairs. These changes in the proportions are statistically significant in both the zebrafish and the amphioxus cases (χ^2 p-values of 3.00 x 10^{-97} and 5.85 x 10^{-5} respectively).

Nevertheless, it is important to stress that although there is a trend to maintain the integrity of some TAD boundaries, that does not mean that genome architecture have remained static. Only 85 out of the 90496 possible pairs of neighbors in zebrafish are conserved from the last common ancestor of deuterostomes, and a similar proportion is found in amphioxus with 38 pairs out of 39530 (Figure 4.12B). Among them it is possible to distinguish some common cases such as the *Smad9/Alg5* syntenic pair that is found in all nine species included in the analysis, which is remarkable. Accordingly, both genes seem to belong to the same RL both in zebrafish and in the european amphioxus when looking at their interactions using HiChIP (Figure 4.12C), as they likely did more than 450 mya in the ancestor of deuterostomes. However, the majority of the pairs are newer and we are particularly interested in those pairs that appeared at the root of vertebrates and are integrated within the same RL. Using our strategy we were able to find a collection of syntenic pairs shared by all five vertebrate species but absent in the rest of species considered. Out of this collection we retained those pairs that contain a gene that projects a RL that extends beyond the new neighbor and obtained a list of 393 candidate pairs. Those candidate pairs might lead us to identify cases of enhancer-promoter rewirings that happened at the root of vertebrates and were important in the invertebrate to vertebrate transition. One of such candidates is the hedgehog signalling receptor *ptch1* and its immediate neighbor *ercc6l2*. They are found together only in vertebrates and the big *ptch1* RL in zebrafish extends largely beyond *ercc6l2* to an entirely

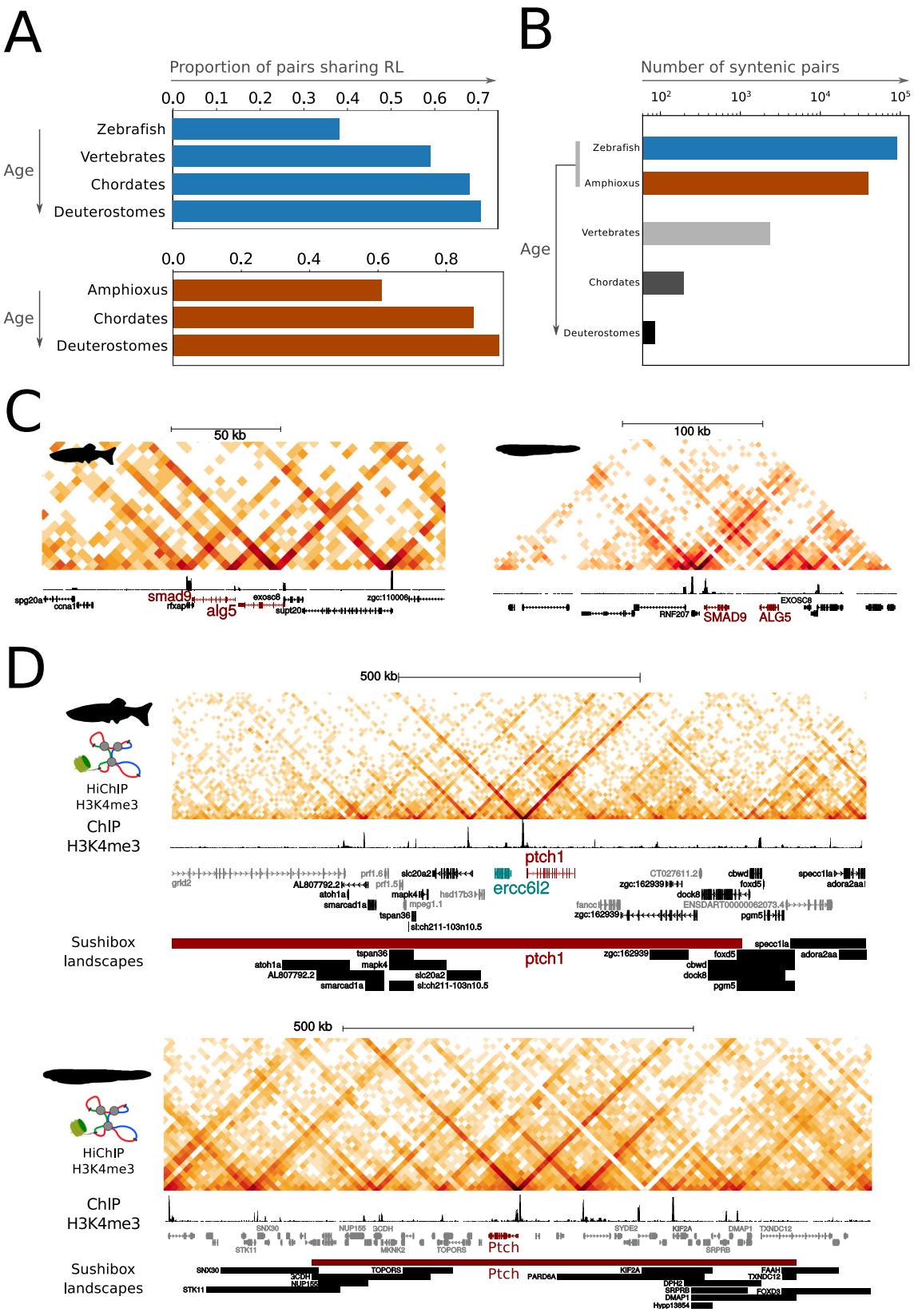


Figure 4.12: Relationship between genomic rearrangements and RL boundaries. (A) Proportion of pairs of neighboring genes that belong to the same RLs according to the age of the pair both in zebrafish (top, blue) and amphioxus (below, bottom). (B) Total number of sytenic pairs per age category. (C) The *Smad9/Alg5* gene pair has remained together at least from the deuterostome ancetor. These two genes share an important part of the landscape both in zebrafish (left) and amphioxus (right). (D) In contrast, the syntenic region located upstream of *ptch1* is only found in the vertebrate lineage. However, it is completely integrated in the RL of the gene.

new genomic region in terms of synteny (Figure 4.12D).

4.2.5

HiChIPs against H3K27ac reveal enhancer hubs around developmental gene promoters

Using HiChIP against H3K4me3 we were able to identify with high resolution contacts involving active promoters both in zebrafish and in amphioxus. Then we wondered if by using the H3K27ac we would be able to recover interactions taking place between enhancers and promoters. Even more, we wondered if with this assay could both identify and assign active enhancers in one go. In order to explore this possibility we performed replicated H3K27ac HiChIP experiments in zebrafish and in amphioxus. For zebrafish we used 80% epiboly and 24 hpf embryos and for amphioxus 15hpf embryos that are equivalent to those used for previously described experiments.

As expected, the interaction matrices obtained were different yet compatible with the interaction matrices of the H3K4me3 HiChIP experiments. Strong diagonal stripes also emerge from active promoters delineating its RL, and the extension of such RLs are equivalent to those obtained with experiments with the H3K4me3 antibody. Apart from those stripes, in the H3K27ac experiments it is also possible to observe additional ones emerging from active enhancers. Often, those lines connect with those belonging to active promoters and in such cases it seems reasonable to assume that we have identified a bona fide enhancer-promoter pair. An example of this can be observed in the Figure 4.13A, that compares the interaction matrices of H3K4me3 and H3K27ac HiChIP experiments around the *znf503* locus in zebrafish. From the H3K4me3 HiChIP experiment it is already possible to observe how the RL of *znf503* extends towards the introns of the neighboring *c13h10orf11* gene. In addition to that, the H3K27ac HiChIP experiment also shows a group of enhancers present in the introns of the neighboring gene interacting frequently both among themselves and with the *znf503* promoter. We termed such associations between a promoter and a numerous group of enhancers *enhancer hubs* and we sought to identify them genome wide in our experiments.

GO-Term	**Count**	**Benjamini p-value**
GO:0006355~regulation of transcription, DNA-templated	109	1.45E-8
GO:0007275~multicellular organism development	61	4.73E-8
GO:0006351~transcription, DNA-templated	63	1.43E-4
GO:0016055~Wnt signaling pathway	20	5.25E-4

Table 4.3: GO associated to enhancer hubs in 80% epiboly zebrafish embryos

GO-Term	**Count**	**Benjamini p-value**
GO:0006355~regulation of transcription, DNA-templated	78	1.22E-6
GO:0007275~multicellular organism development	39	4.74E-4
GO:0006351~transcription, DNA-templated	46	8.15E-4
GO:0007420~brain development	16	0.03

Table 4.4: GO associated to enhancer hubs in 24hpf zebrafish embryos

We followed the rationale that the signal distribution around enhancer hubs resembled those of isolated TADs and therefore we calculated insulation scores. We found that unusually high

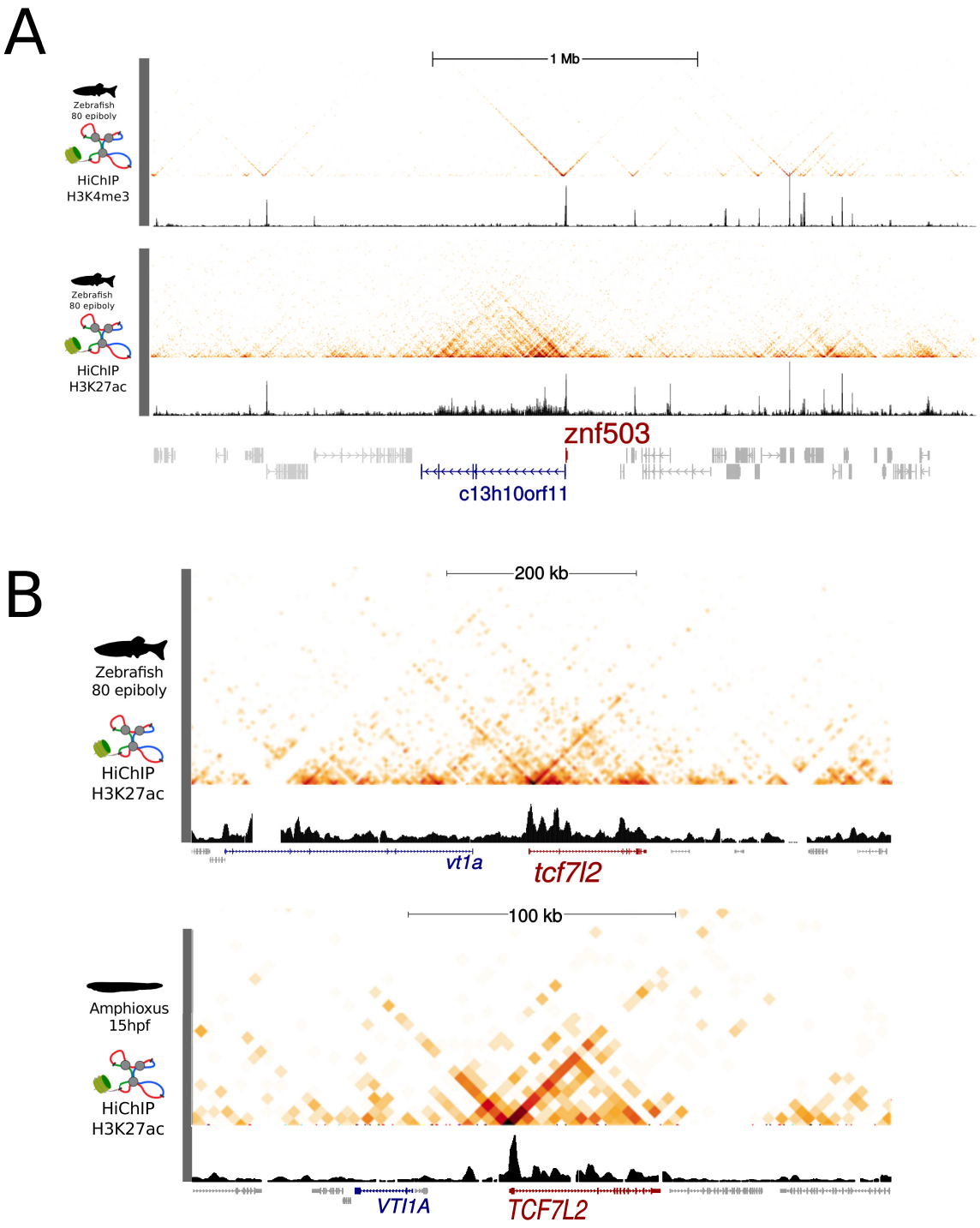


Figure 4.13: Enhancer-promoter hubs are frequently associated with developmental promoters. (A) Comparison between the H3K4me3 (top) and the H3K27ac (bottom) HiChIP experiments in zebrafish around the *znf503* locus. HiChIP derived ChIP-seq tracks are shown below the contact matrices. Additional stripes emerge from active enhancers connecting with promoters in the H3K27ac experiment. (B) Conserved enhancer-promoter hub involving the *Tcf7l2* gene and the *Vt1a* syntenic gene between zebrafish (top) and amphioxus (bottom).

values of insulation scores were highly predictive of the presence of an enhancer hub (details and code in the Material and Methods section 3.2.2). We were able to find 333 and 483 enhancer hubs in the zebrafish HiChIPs performed in 24hpf and 80% epiboly embryos respectively. Probably due to problems with the contiguity in the assembly of amphioxus, which hinders the calculation of insulation scores, we were only able to find 45 of such enhancer hubs in this species. Once calculated, we decided to check the function of the genes associated to them and performed Gene Ontology Enrichment Analysis using DAVID in both zebrafish experiments. As perhaps could be expected, the most enriched terms were related with transcriptional regulation and developmental processes (Tables 4.3 and 4.4). Then, genes involved in such structures seem to be tightly regulated, maybe due to a high hierarchical position within their GRNs. One of such genes is *Tcf7l2*, a well known TF belonging to the Wnt signalling pathway (Figure 4.13B). We found it associated to a well conserved enhancer hub that can be identified both in zebrafish and in amphioxus. Those two hubs are probably homologous because they encompass an equivalent syntenic region that includes the introns of both *Tcf7l2* itself and the conserved neighbor *vt1a*.

4.2.6

Polycomb mediated long range interactions between developmental promoters in vertebrates

So far we have only explored active epigenetic marks, and therefore the contacts that are established by promoter and enhancer regions in those cell populations where they are turned on. We then wondered if there were differences in the contacts established by promoters when they are active in comparison to when they are inactive. To explore this further we performed HiChIP against the H3K27me3 histone modification, which is deposited by the PRC2 Polycomb related complex and is related to the facultative repression of transcription. We performed replicated experiments both in zebrafish (again equivalently staged 24hpf embryos) and in amphioxus (15hpf embryos) and we first inspected visually the interaction matrices obtained.

In zebrafish, long range contacts connecting distant promoters that are decorated with the H3K27me3 are readily observed. Strikingly, such contacts cross the boundaries of the RLs that can be defined using H3K4me3 HiChIP experiments as can be appreciated in Figure 4.14A. In particular, it is possible to observe how the hoxD cluster, the *sp* genes and the *dlx1a/dlx2a* pair of genes are all specifically decorated with the H3K27me3 mark and contacting each other. Owing to the fact that we are using whole embryos this can be interpreted as follows: in some cell populations of the 24hpf zebrafish embryo those promoters are specifically repressed by the PRC2 complex and, furthermore, is in this context where they tend to interact with each other. Meanwhile, in cell populations where these genes are active they interact less frequently among themselves.

In contrast, such far range contacts between Polycomb repressed promoters are less obvious in amphioxus H3K27me3 HiChIPs, although it is difficult to rule them out completely (Figure 4.14B). It is worth noting that the H3K27me3 signal is also sparser in amphioxus, although clear enrichments are visible around specific areas such as the RL of the Hox cluster or the area surrounding the *Gbx* gene. In order to have a clearer view on how these contacts evolved we decided to also inspect publicly available H3K27me3 data from the KC167 cell line of *Drosophila melanogaster* (Rowley et al. 2017). In *Drosophila*, H3K27me3 signal is more evenly distributed than in zebrafish, covering entire RLs rather than being localized sharply around promoters. However, far range

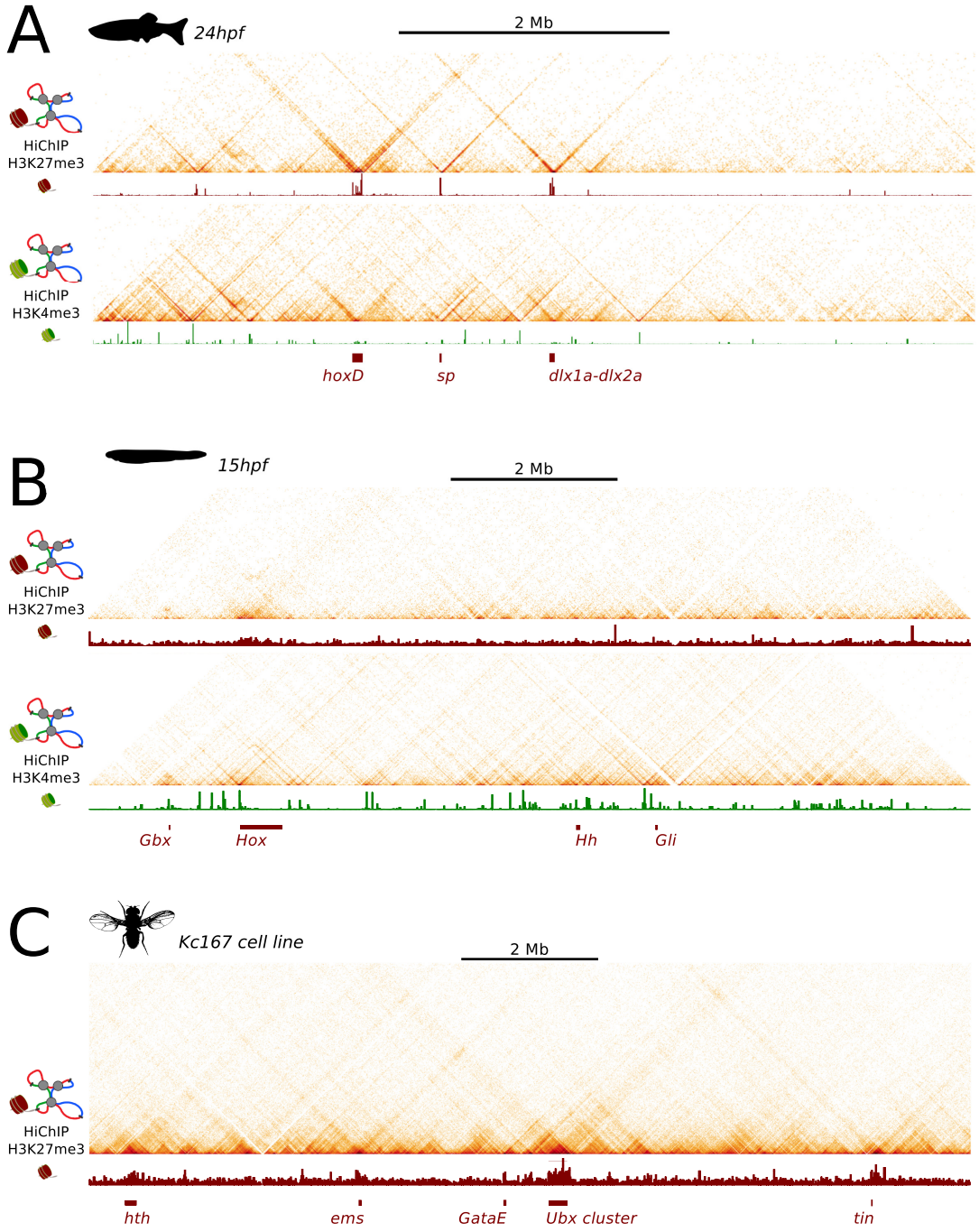


Figure 4.14: Vertebrate developmental promoters interact while they are repressed by the PRC2 complex. (A) Comparison between H3K27me3 (top) and H3K4me3 (bottom) HiChIP experiments in zebrafish 24hpf embryos. Only the names of some Polycomb enriched promoters are shown. Below the contact matrices the HiChIP derived ChIP-seq signal is also shown. (B) Equivalent comparison to the one in (A) but for 15hpf amphioxus embryos. (C) For the *Drosophila* Kc167 cell line only the H3K27me3 HiChIP is available (Rowley et al. 2017).

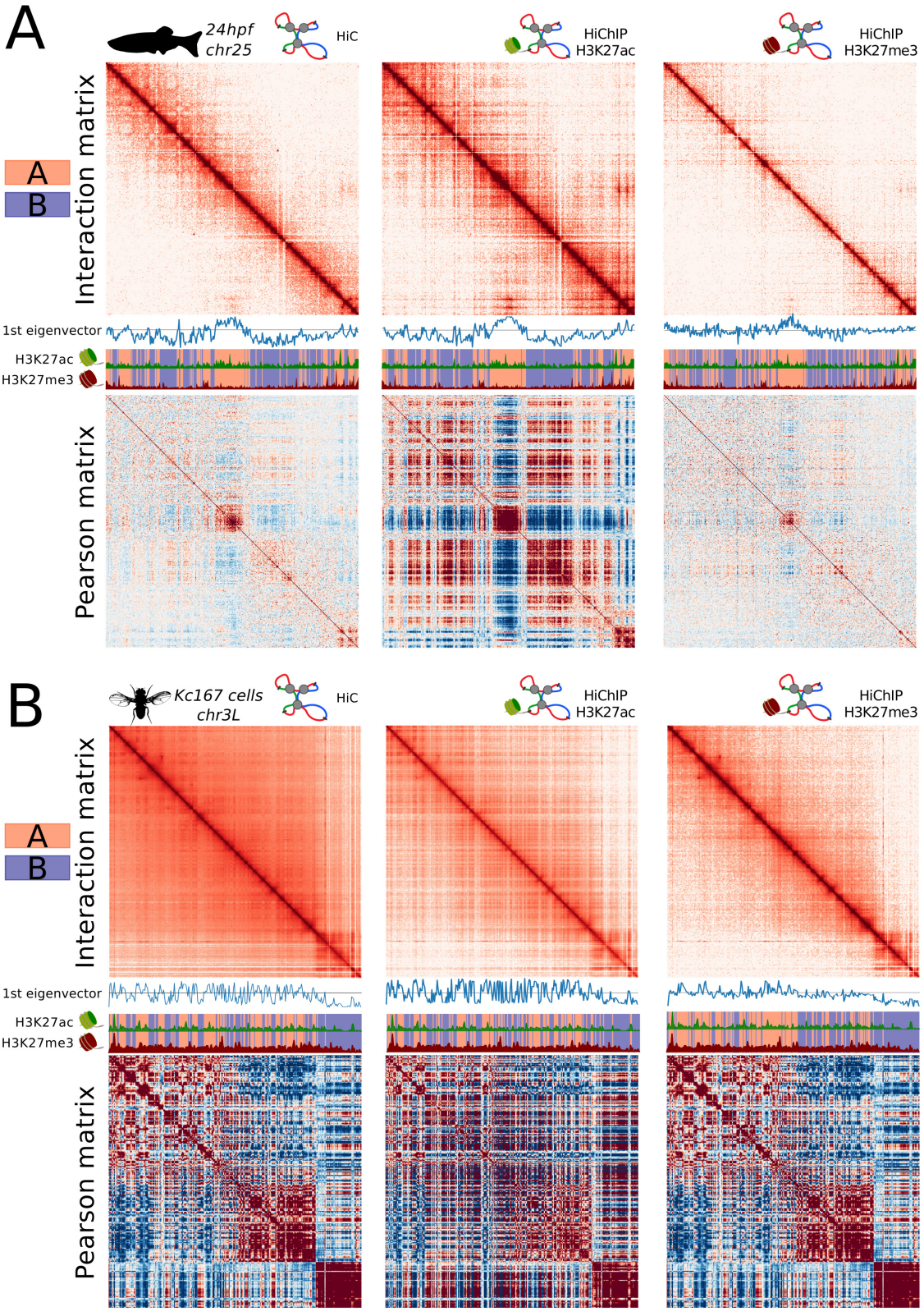
contact between whole RLs decorated with H3K27me3 can be easily spotted, for instance between the *hth* and the *ems* RLs or between the *Ubx* cluster and *tin* (Figure 4.14C). Given that such long range contacts observed mainly in flies and zebrafish clearly overflowed the limits of TADs and RLs, we decided to check their relationship with the next classical hierarchy of nuclear interactions: the A and B compartments.

In order to see how those far range interactions were related to compartments we used available HiC experiments both in zebrafish and in flies together with H3K27ac and H3K27me3 HiChIP experiments in those two species and in amphioxus. We calculated interaction matrices at 100kb resolution for all three kinds of experiments. Then, we calculated the pearson correlation transformation of such matrices and the first eigenvector, which is the usual procedure to unravel compartments A and B from HiC experiments (4.15, details and code in Material and Methods section 3.2.2). We also calculated smoothened H3K27me3 and H3K27ac signals to 100kb windows.

In zebrafish, A and B compartments could be easily identified from the pearson matrices from both the HiC and the H3K27ac HiChIP experiments, as it can be observed in Figure 4.15A. Both of them displayed equivalent chessboard patterns and similar trends in the eigenvector values, which resulted in very similar compartment calling (see A/B compartments in the introduction). As expected, H3K27ac occupancy correlated well with one of the two compartments which is subsequently considered the active compartment or compartment A. The H3K27me3 HiChIP pearson matrix was blurrier than the other two and the compartment calling was the most divergent although not entirely incompatible. This is possibly ought to the unequal coverage resulting from selecting only contacts involving H3K27me3 regions, which are small and sparsely distributed in zebrafish as seen in Figure 4.15A. This problem does not exist neither when calculating compartments from HiCs nor in the case of the H3K27ac HiChIPs since this mark decorates the A compartment pervasively. In fact, the compartment signal is even sharper in the second case. Intriguingly, H3K27me3 enriched areas seem to belong to the A compartment in zebrafish, although it could be argued that many of them are located on the edges with the B compartment. This would be, in principle, at odds with the classic vision of repressed chromatin domains belonging to the B compartment.

Similarly, sharp compartment signals can be extracted from both the HiC and the H3K27ac HiChIP matrices in *Drosophila* (Figure 4.15B). Expectedly, much sharper compartment signal also emerged from the H3K27me3 HiChIP experiment, probably due to the less localized distribution of the H3K27me3 signal throughout the *Drosophila* chromatin. However, it is worth noting that both in zebrafish and *Drosophila* the most divergent compartment calling was always the one calculated from the H3K27me3 experiment. Nevertheless, in stark contrast with zebrafish, H3K27me3 enriched regions are mostly located inside the B compartment. Furthermore, the H3K27me3 signal seems to anticorrelate with both the H3K27ac signal and the first eigenvector values. Finally, similarly to the case of *Drosophila*, it is also possible to appreciate different compartments in the chromosomes of amphioxus embryos using both types of HiChIP experiments (Figure 4.15C). In addition, the H3K27me3 mark also seems to slightly anticorrelate with the H3K27ac mark.

Shockingly, although we still lack important pieces of information, it is tantalizing to speculate that in vertebrates Polycomb repressed chromatin domains behave in a different way. Somehow, they seem rather disconnected from the classic dichotomy between the A and the B compartment. This topic will be further discussed in following sections.


A
24hpf
chr25
HiC
HiChIP
H3K27ac
HiChIP
H3K27me3
Interaction matrix
A
B
1st eigenvector
H3K27ac
H3K27me3
Pearson matrix
B
Kc167 cells
chr3L
HiC
HiChIP
H3K27ac
HiChIP
H3K27me3
Interaction matrix
A
B
1st eigenvector
H3K27ac
H3K27me3
Pearson matrix

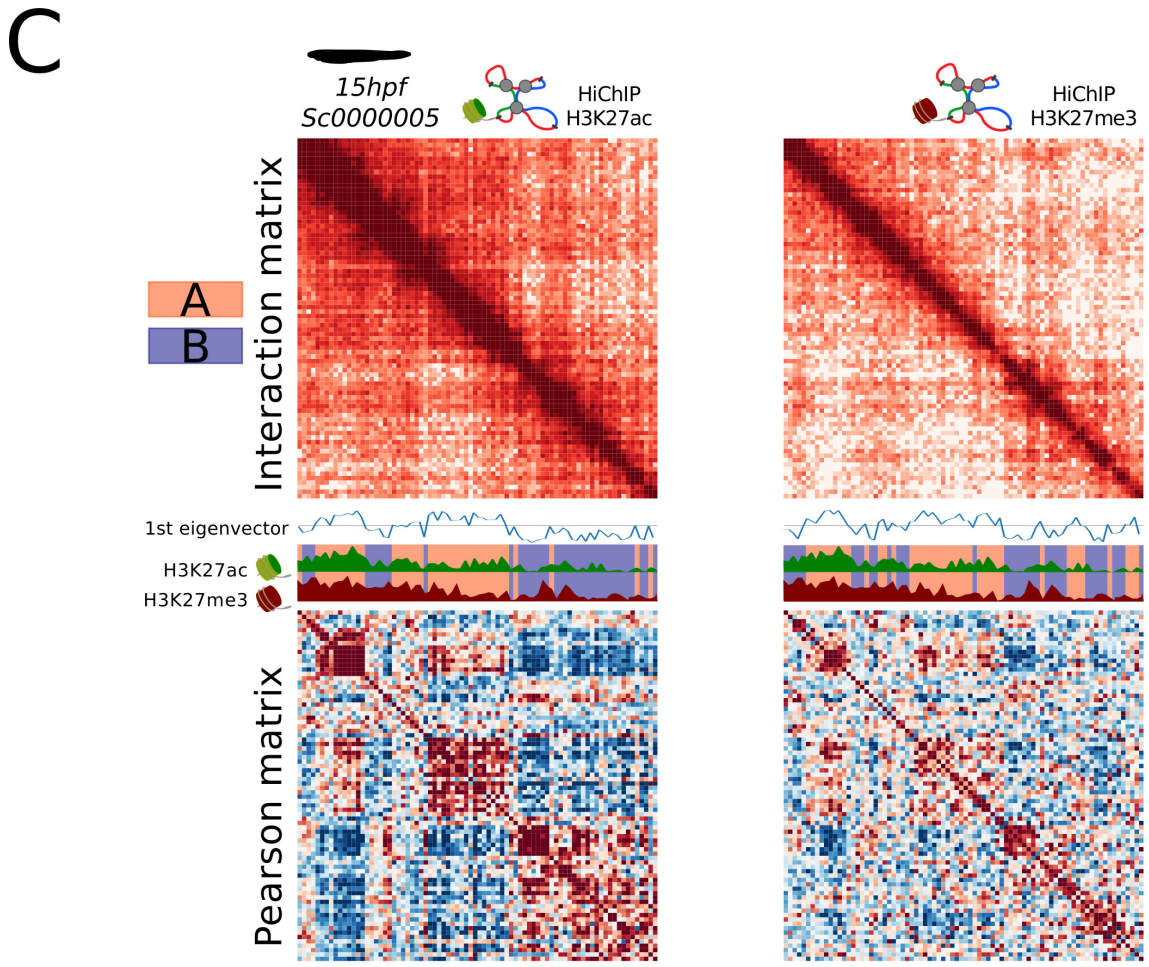


Figure 4.15: Relationship between long range contacts between repressed promoters and compartments. (A) Interaction matrices (top) and pearson corrected matrices (bottom) from zebrafish HiC (Kaaij et al. 2018) and zebrafish H3K27ac and H3K27me3 HiChIP experiments at 100kb resolution. The first eigenvector or principal component is shown in the middle together with the HiChIP derived ChIP-seq signal for H3K27ac/H3K27me3 also binarized to 100kb. The predicted A compartments are colored orange and the predicted B compartments are colored blue. (B) Same plots than in (A) but for the Kc167 *Drosophila* cell line (Rowley et al. 2017). (C) Same plots for amphioxus embryos, but no HiC experiment is available.

Chapter 5

Discussion

5.1

Are TADs a synapomorphy of animals?

So far we have presented results that compare the folding of the genomes between vertebrates and cephalochordates, both genome wide and at several specific loci. However, genome folding can be compared at multiple scales, from chromosome territories to individual loops. In this work we have focused on how chromatin folds at the submegabase scale, which is the one that is more relevant for the specificity of the interactions between enhancer and promoters. At this particular scale, the compartmentalization of the genome in TADs is the most prominent feature and has been identified across animals of different phyla. In short, this kind of folding facilitates the interactions between enhancers and promoters belonging to the same TAD, while preventing contacts between enhancers and promoters of different TADs. On one hand, in our comparison we have explored how the size of TADs evolved, since bigger TADs can potentially accommodate more enhancers. On the other, we have pinpointed places where genomic rearrangements involving TAD boundaries could have generated new enhancer-promoter contacts. As we will discuss later, by doing this we have found both individual events and general trends that were needed for the appearance of several important regulatory novelties in the vertebrate lineage. However, it is important to note that many of the conclusions that can be extracted regarding how the folding of genomes evolves can only be generalized to the cases of organisms with TADs. Therefore, the question of when did chromatin organization in TADs first evolve and how conserved this type of organization is across different taxa is central for us.

Our knowledge of the presence or absence of TADs in the different animal species is starting to rapidly grow but is still fragmentary (Figure 5.1). For instance, chromatin folding in mammals has been extensively profiled with HiC, with contact maps from more than 50 species currently available (the DNA zoo project: Dudchenko et al. 2017). All of them display TADs and the location of the majority of TAD boundaries are preserved (Vietri Rudan et al. 2015). However, outside mammals, the taxon sampling is much scarcer. In other vertebrates we only count with HiC experiments from zebrafish (Kaaij et al. 2018) and chicken (Gibcus et al. 2018), together with several 4C-seq experiments from medaka (Letelier et al. 2018a) and snakes (Guerreiro et al. 2016). All of them either directly prove (when HiC data is available) or strongly indicate the presence of TADs in each

of the species. In this work we report that the european amphioxus (*Branchiostoma lanceolatum*) also displays TADs by using two different approaches. On one hand we performed arrays of 4C-seq experiments followed by computational modelling in the Hox locus, finding how the Hox cluster is completely embedded within a single TAD. On the other we performed replicated HiChIP experiments targeting three different epigenetic marks (H3K4me3, H3K27ac and H3K27me3) and compartmentalized RLs are also readily observed genome wide. The observation of a conserved TAD boundary bisecting the Six cluster of sea urchin (Gómez-Marín et al. 2015), as inferred from 4C-seq experiments, also indicate that TADs were probably present in the deuterostome ancestor. Meanwhile, in protostomes, TADs have been identified using HiC both in fruit flies (Sexton et al. 2012) and mosquitoes (*Aedes aegyptii*, Dudchenko et al. 2017) and have been found to be absent in the nematode *Caenorhabditis elegans*. The absence of TADs in *C. elegans* might be related to the secondary loss of the architectural protein CTCF specifically in this particular nematode lineage. Here we also provided proof for the presence of TADs in the centipede *Strigamia maritima* by investigating the chromatin architecture of the Hox locus using the 4C-seq coupled to modelling strategy. The finding of TADs in myriapods, together with the presence of TADs in flies and mosquitoes, further indicate that they were also likely present in the common ancestor of arthropods. Unfortunately, we currently lack chromosome conformation capture experiments from both spiralians and non bilaterian groups such as cnidarians, sponges, ctenophores or placozoans. In any case, despite the absence of architectural data in several key groups, it seems reasonable to assume that TADs appeared early in the evolutionary history of animals and most probably were already operating at least in the last common ancestor of bilaterians. Apart from the fact that TAD structures can be found both in the protostome and the deuterostome lineages, there are additional although indirect lines of evidence that indicate that the bilaterian ancestor had TADs.

First of all, the architectural protein CTCF is found in most bilaterian phyla but not outside this group (Heger et al. 2012). CTCF loss of function experiments in human cells lead to the complete dismantling of TAD boundaries (Nora et al. 2017). Therefore, the appearance of this protein at the origin of bilaterians might have been the decisive factor enabling the building of TADs. However, it is worth noting that the role of CTCF in the protostome lineage is still not fully established. In deuterostomes, CTCF proteins convergently bound to the DNA are able to dimerize and stop cohesin rings stabilizing chromatin loops (Sanborn et al. 2015). Accordingly, TAD boundaries are enriched for arrays of divergently oriented CTCF binding sites that provide insulation, a pattern that is consistent from mammals (Rao et al. 2014) to echinoderms (Gómez-Marín et al. 2015). Strikingly, this does not seem to be the case in flies where no such divergent pattern is found at TAD boundaries where CTCF is bound (Rowley et al. 2017). Furthermore, in *Drosophila*, actively transcribed regions appear to be better predictors of boundaries than architectural proteins binding sites including CTCF. Such active regions will belong to the A compartment, and the presence of small A compartments in between two bigger B compartments is enough to generate insulation. Then, additional experiments exploring CTCF occupancy and orientation at TAD boundaries in other protostomes are needed in order to shed more light on the mechanisms of TAD formation in the bilaterian ancestor.

Secondly, animals display a quite disproportioned amount of examples of deep conservation of microsynteny when compared to other eukaryotes (Irimia et al. 2012). Such conservation has been traditionally linked to the presence of distal enhancers that regulate genes that are not strictly the nearest. Then, genomic rearrangements breaking synteny at such places will disconnect enhancers and target promoters and that would be disfavored. In the actual paradigm it is difficult to envision

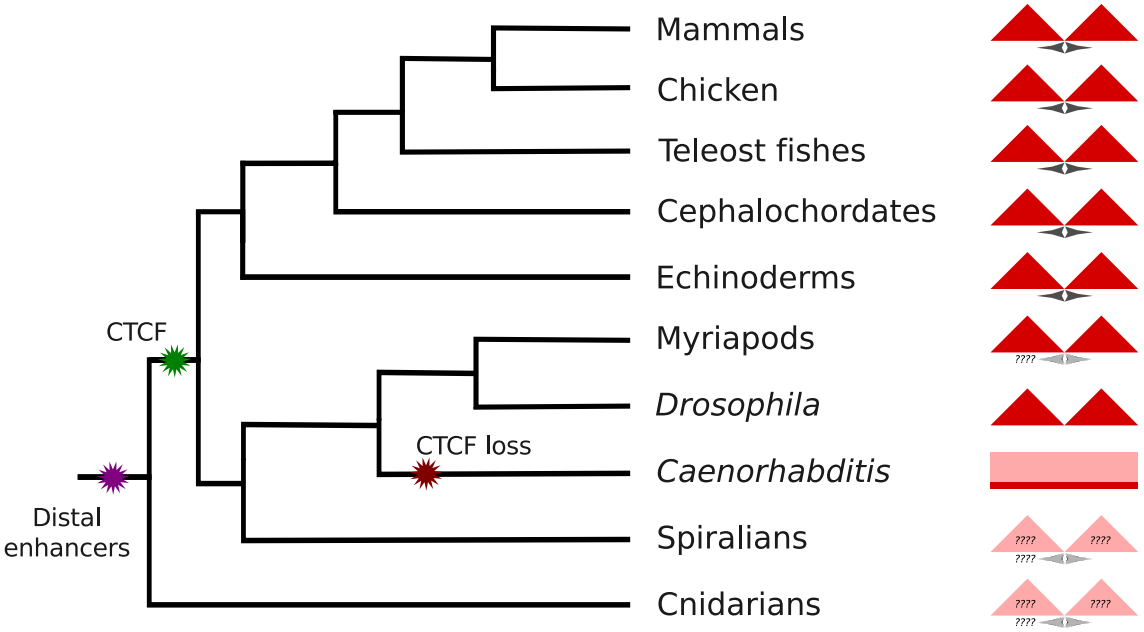


Figure 5.1: Phylogenetic tree depicting some known and unknown cases of the presence of TADs. The appearance of distal enhancers before the split of bilaterians and cnidarians, the appeareance of CTCF at the root of bilaterians and the loss of CTCF in the *C. elegans* lineage are also shown. The presence of TADs is marked with dark red triangles and the mechanism of formation based on diverging CTCF sites at TAD boundaries is depicted with black arrows. Pink triangles and grey arrows mean unknown presence or absence of TADs and unknown TAD formation mechanism respectively.

how regulation through distal enhancers would happen in the absence of TADs, and therefore the constraints to maintain microsynteny might indicate the presence of such structures. Indeed, syntenic blocks and gene regulatory blocks are good predictors of the location of TAD boundaries (Harmston et al. 2017,Krefting, Andrade-Navarro, and Ibn-Salem 2018). This probably reflects that the pressure to not break TADs is an important force in the special preservation of microsynteny in animals. In concordance, we also observe this trend in our H3K4me3 HiChIP experiments both in zebrafish and in amphioxus. In our case, we found that the longer time a pair of genes have remained syntenic, the most likely is that they are encompassed by the same RL. This probability reaches a 90% in the case of syntenic pairs conserved from the root of deuterostomes. Accordingly, the number of syntenic pairs that the nematode *C. elegans* share with other animals is extremely low (only 12, Irimia et al. 2012), which might be a consequence of the TAD losses in this lineage.

Finally, the question of whether TADs can be found or not outside bilaterians remains open and of special interest is the case of cnidarians. *Nematostella vectensis*, for instance, displays a remarkably high number of syntenic pairs shared with other bilaterians (200), despite the fact that all cnidarians lack CTCF (Irimia et al. 2012). This number is in the same order of magnitude than the number of pairs retained by the putative vertebrate ancestor (374) and clearly higher than the number of pairs kept by *Drosophila melanogaster* (46). In addition, the H3K4me1 epigenetic modification that is a signature of active distal enhancers operates in *Nematostella vectensis* equivalently to other bilaterians. Indeed, transgenic reporter assays performed in this cnidarian species reveal the enhancer activity of distal genomic elements. Taking these elements together with the fact that *Drosophila melanogaster* displays TADs in a seemingly CTCF independent manner, it is conceivable to speculate that TADs originated before CTCF. In contrast, the origin of cohesin

rings, which are perhaps the other best characterized player in the formation of TADs, is very ancient. In any case, chromatin conformation capture experiments in additional protostomes and non-bilaterians are needed to further clarify when and how TADs originated. Such experiments have been performed outside metazoans, for instance in other eukaryotes like the yeast species *Saccharomyces cerevisiae* (Duan et al. 2010) and *Schizosaccharomyces pombe* (Mizuguchi et al. 2014), the protozoan *Plasmodium falciparum* (Ay et al. 2014) and the plant model *Arabidopsis thaliana* (Wang et al. 2015). Intriguingly, in *S. pombe*, chromatin compartmentalization in globules ranging from 50kb to 100kb has been described and these structures form in a cohesin dependent manner (Mizuguchi et al. 2014). Therefore, they would be similar in size to the TADs described in *Drosophila*, but no relationship between such chromatin arrangements and the control of gene regulation has been yet established. It is worth noting, however, that the genome of *Drosophila* is much larger and intergenic distances are wider than in *S. pombe*. A similar scenario is found in the only chromosome of the bacteria *Caulobacter crescentus* (Le et al. 2013), with equivalently sized chromatin compartments named CIDs (for Chromatin Interacting Domains). It is conceivable, though, that resolution is an issue in order to find a functional equivalent to TADs in organisms with smaller genomes. Traditional C-techniques are limited in resolution by the cut frequency of restriction enzymes and perhaps the further development of protocols like Micro-C (Hsieh et al. 2015), that reaches nucleosomal resolution, will shed some light on this discussion.

5.2

The evolution of cis-regulatory elements in the context of TADs

If we assume that TADs are both critical to enable enhancer-promoter contacts and also a widespread feature of the bilaterian genomes, then it is important to bear them in mind when discussing questions related to the evolution of transcriptional regulation in animals. Perhaps one obvious implication of the emergence of TADs is that it allowed regulatory regions to find and regulate promoters that are located far away in the linear sequence. This is important if we consider the problem of the emergence and wiring of new enhancers. There are a number of processes by which a new enhancer can appear including (i) de novo generation through mutations that generate TF binding sites, (ii) modification of previous enhancers, (iii) duplication and modification of previous enhancers and (iv) the insertion of transposable elements carrying enhancers or sequences that can be rewritten to enhancers (as proposed in Maeso and Tena 2016). It is important to note that enhancers are complex entities that, apart from containing TF binding sites, they need to be able to open the surrounding chromatin and make themselves visible to the machinery of transcription. Then, producing new enhancers does not seem to be an easy process from an evolutionary point of view. Furthermore, such processes would be hindered even more if they needed to occur precisely in the vicinity of the gene that is going to be regulated. Related to that, in our comparative 4C-seq experiments between zebrafish and mouse we observed how TADs seem to be flexible to changes in size while maintaining their boundaries in terms of synteny. From that it can be interpreted that mutational events including insertions and deletions are generally permitted if they do not perturb the boundary elements. In that sense mechanisms such as the duplication of enhancers

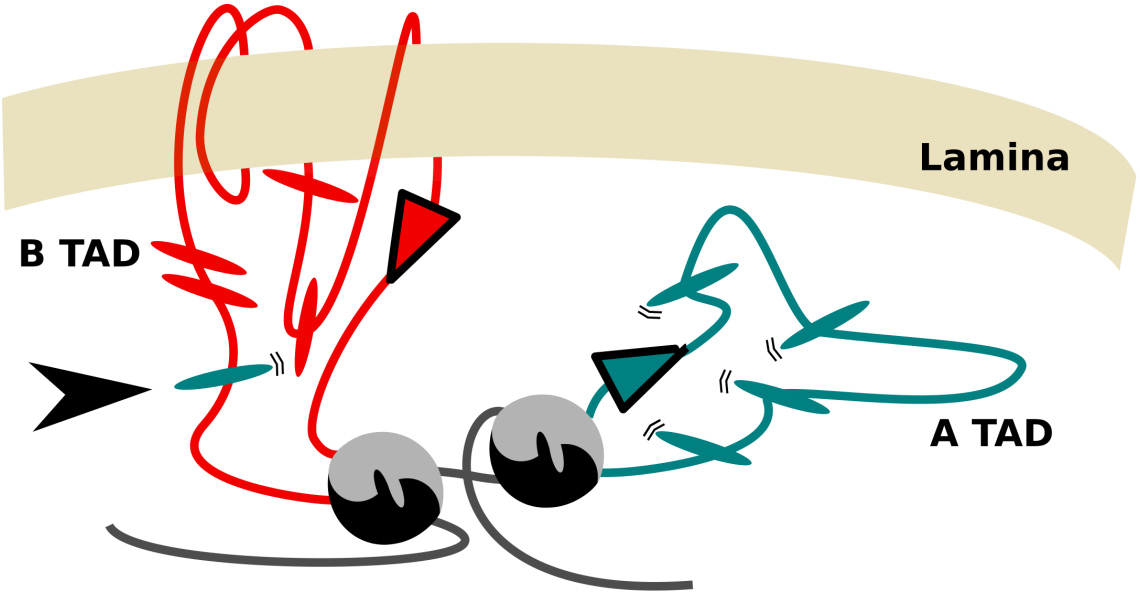


Figure 5.2: The appearance of a single active enhancer in a new TAD (blue oval marked with an arrowhead in the B TAD) might not be sufficient to activate a target gene if the enhancer gets activated in a completely different context than the rest of enhancers of the TAD. Enhancers are ovals, promoters are triangles. The boundary between the TADs is shown with CTCF locks.

or the insertion of transposable elements, for instance, are allowed in the context of TADs and in many cases probably contributed to the expansion of both the size and the regulatory complexity of some of them. Nevertheless, although in principle most enhancer-promoter contacts happening within a TAD are thought to be favored, the constraints imposed by internal architecture of TADs in the incorporation of new enhancers need to be fully addressed. In summary, it is plausible that the appearance of TADs helped to widen importantly the range of locations where regulatory elements can appear and function. Indeed, it is tempting to speculate that the former could have contributed to the burst of complexity in the temporal and spatial control of gene expression in the animal lineage (Heger et al. 2012).

On the other hand, it can also be argued that TADs reduce the pleiotropic effects of incorporating new regulatory elements. It is well established that enhancers have the potential to regulate more than one gene, and that is the reason why certain clusters of developmental genes like the Hox genes (Duboule and Dollé 1989) or the Irx genes (Tena et al. 2011) tend to remain together in many organisms. By confining the regulatory potential of newly evolved enhancers to the promoters belonging to the same TAD, potentially harmful pleiotropic effects are diminished. Furthermore, it is important to note that in many cases TADs have been shown to behave as coordinated units. For instance TADs tend to switch from active to inactive compartments as a whole (Dixon et al. 2015), and equivalently they also tend to associate to lamina in a coordinated manner. Then, the odds of a newly evolved enhancer of making an impact in the transcriptional levels of a given gene might be higher if it drives expression to similar tissues than their partners that operate in the same TAD (Figure 5.2). This might explain why redundant and shadow enhancers are commonly found (Cannavò et al. 2016), and it may also mean that quantitative changes in gene expression and the refinement of expression domains are much more common than the gain of new expression territories in an utterly different context. In contrast, genome compartmentalization in TADs provide and additional and more drastic mechanism to produce the gain of new expression domains.

That would be altering the TADs in a way that a full set of enhancers are able to interact with a promoter that was never accessible before. We will explore this further in the next section.

5.3

Are changes in the 3D topology a relevant mechanism in the evolution of GRNs?

A number of studies have shown that altering the physiological nuclear architecture, specifically at the level of the formation of TADs, is sufficient to produce relevant modifications in gene expression. For instance, upon the depletion of CTCF (Nora et al. 2017) or cohesin (Rao et al. 2017), important changes in the transcriptome occur. However, from an evolutionary angle, the lost or modification of these proteins and the general dismantling of the chromatin architecture would most often produce deleterious effects. Indeed, CTCF and cohesin are both essential for mice embryonic development (Wan et al. 2008, Schwarzer et al. 2017) and they are conserved throughout the animal kingdom with very little exceptions. This is expected, since TADs have been shown to be essential in order to connect most developmental regulators with their set of enhancers (Symmons et al. 2016). Mirroring the classical prediction stating that enhancers and CREs are more evolvable than TFs, TAD boundaries and its their position are also conceivably more evolvable than the architectural proteins involved in generating all of them.

There are several ways in which the TAD boundary content of a genome can evolve so that it affects gene expression, including the following: (I) a TAD boundary can disappear, through a deletion or through point mutations on CTCF binding sites, with enhancers and promoters laying on different sides of the boundary becoming then accessible; (II) a new TAD boundary can appear, either de novo or more likely it may come included into a transposable element or be the result of a tandem duplication, then separating a promoter from some of its former enhancers; (III) a structural variant or a genomic rearrangement (e.g. an inversion) may alter the relative position between a set of enhancers, a promoter and a TAD boundary either connecting them the novo or separating them. Indeed, several cases of the third type have been described linking structural variants with altered gene expression and disease. For instance, different rearrangements encompassing the boundaries between the three TADs that harbor the genes *WNT6* and *IHH, EPHA4* and *PAX3* respectively have been related to limb malformations (Lupiañez et al. 2015, Figure 5.3A). Such malformations are caused by the ectopic expression of either *WNT6, IHH* or *PAX3* in limbs, that happens upon the connection of their promoters to limb enhancers present in the *EPHA4* TAD. In the same direction, large duplications including boundary elements can produce new TADs with duplicated genes exposed to different enhancers. One of such events, that duplicate the *KCNJ2* gene and expose it to a group of *SOX9* limb enhancers, provokes the limb abnormalities associated to the Cook's syndrome (Franke et al. 2016). Interestingly, this process leaves one *KCNJ2* copy intact with the old regulatory inputs and a new one with the limb enhancers. In addition, it is also remarkable that 26% of the recurrent small mutations found in T-cell acute lymphoblastic leukemia overlap TAD boundaries and some of them are able to cause the overexpression of protooncogenes like *LMO2* or *TAL1* by connecting them with ectopic enhancers (Hnisz et al. 2016). Finally also translocations, such as the one merging the RLs of *FOXO1* and

PAX3, are able to generate aberrant gene expression and tumorogenesis due to enhancer promoter rewirings (Vicente-García et al. 2017).

Those four examples are suggestive from an evolutionary perspective, since all of them trigger phenotypes associated with the gain of function of a gene in a new context. Indeed, depending on the hierarchical position of that gene in a developmental GRN, altering the position of a TAD boundary could potentially activate an important collection of downstream genes to an entirely new location. Such redeployments of fractions of developmental GRNs to new contexts have been reported often, but they were traditionally linked to the gain of new enhancers rather than to the rewiring of preexisting ones. The *SOX9-KCNJ2* case perhaps has an additional interest because it links in a single mutational event the duplication of a gene and its possible neofunctionalization due to its expression in a new domain. The example of the leukemia-associated recurrent mutations affecting boundaries is also tempting because such reorganizations appear to be under positive selection. Nevertheless, they are selected in the special context of a malignancy and that is likely not related to the probability to be fixed in a species. In fact, studies comparing HiC between four different mammals showed no structural variant altering TAD boundaries between them (Vietri Rudan et al. 2015). Rather, the genomic rearrangements that were spotted were precisely located at TAD boundaries. This tendency has been recently shown to hold from mammals to teleosts, with synteny breaks strongly enriched at TAD boundaries (Krefting, Andrade-Navarro, and Ibn-Salem 2018). Such a pattern seem to be driven by two forces: on one hand the rewiring of enhancers has many potential deleterious effects, on the other hand double strand breaks generated by the TOP2B topoisomerase are more common in TAD boundaries due to a higher topological stress localized there (Canela et al. 2017). However, the catalog of species with genome wide chromosome conformation capture data is still sparse and unevenly distributed, and the importance of structural variants in the evolution of gene regulation remains to be fully determined. Then, expanding this catalog of species and adding species that are more distant (e.g. allowing to compare mammals with other vertebrates or even other deuterostomes) seem to be necessary to clarify this question. Indeed, manual evaluation of synteny coupled to individual 4C-seq experiments comparing developmental gene RLs between vertebrates and cephalochordates was sufficient to spot some of such cases (e.g. the evolution of the *Otx* RL, see Figure 4.9).

In that regard, the recent release of high quality genome assemblies combined with HiC data from more than 50 mammalian species will be of paramount interest (the DNA zoo projet, Dudchenko et al. 2017). In our case, we chose the strategy of generating HiChIP data from both zebrafish and amphioxus embryos in order to look for genomic rearrangements that could have contributed to regulatory changes in the origin of vertebrates. Remarkably, we found that the pressure to maintain the integrity of TADs and RLs endured in some cases more than 400 million years of separate evolution between zebrafish and amphioxus. That is the case of the *Smad9* and *Alg5* genes, kept together by the same RL at least from the last common ancestor of chordates (see Figure 4.12), as occured with at least another 231 pair of genes. However, in contrast to what happened when comparing different mammalian species, the synteny of many RLs has been reshuffled when comparing cephalochordates with vertebrates. We were particularly interested in examples of vertebrate specific syntenic blocks included within the same RL in zebrafish, because they could have been relevant for the evolution of new vertebrate traits. We found 393 of such blocks, which encourages us to keep exploring the possibility that TAD reorganizations played an important role in the origin of vertebrate regulatory novelties.

However, it is worth noting that zebrafish and amphioxus are very distant phylogenetically

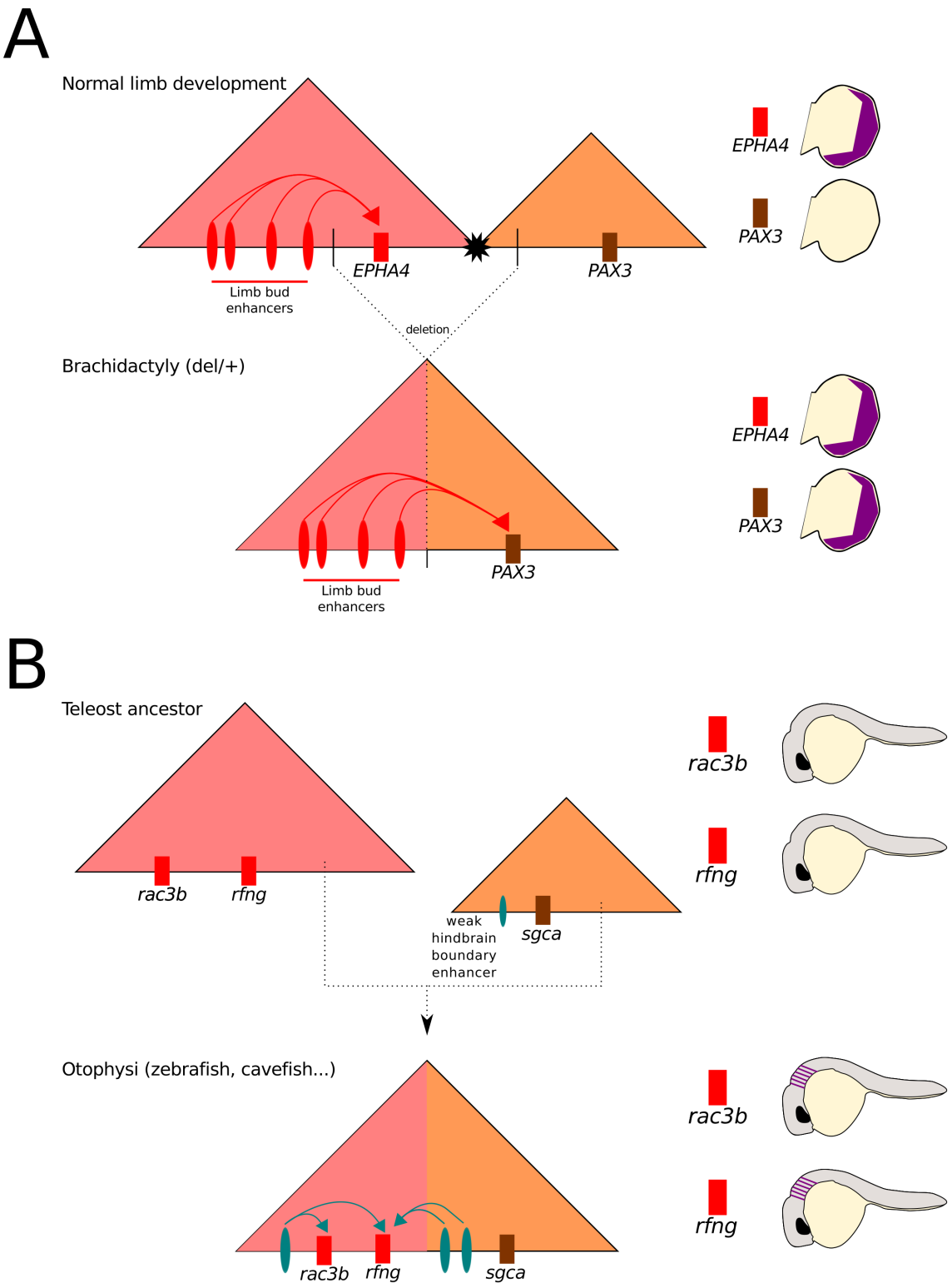


Figure 5.3: Two examples of changes in gene regulation produced by changes in chromatin architecture. (A) From Lupiañez et al. 2015. In normal human limb buds only the *Epha4* enhancer is expressed thanks to the limb bud enhancers present in the red TAD. However, if there is a big deletion in one of the alleles encompassing the TAD boundary between the *EPHA4* TAD and the *PAX3* TAD the *PAX3* gene might contact former *EPHA4* limb enhancers and become expressed in limb buds generating Brachidactyly. (B) From Letelier et al. 2018b. A genomic rearrangement specific from the *Otophysi* lineage connected the TAD containing the *rac3b* and the *rfng* genes with the TAD of *sgca*. The latter brought with him a weak hindbrain boundary enhancer that expanded in this lineage and allowed the *rac3b* and the *rfng* genes to be expressed in that domain. The expression of *rac3b* in the hindbrain boundaries is essential in order to build actomyosin cables in the boundaries, a novelty of this lineage.

and that means that it would be always difficult to assess if such structural variants immediately generated regulatory innovations or, rather, just provided a substratum for further elaboration. In fact, perhaps two of the specific loci where evolutionary changes in the 3D topology have been most carefully studied are examples of the second scenario. The first of them involved the addition of the *sgca* gene to *rac3b/rfng* syntenic block specifically in the *Ostariophysi* lineage, which include species like zebrafish and cavefish (Figure 5.3B). Both in zebrafish and in cavefish, such rearrangement is accompanied by the expression of those three genes specifically in the hindbrain boundaries, expression that is not found outside this lineage (Letelier et al. 2018b). The expression of *rac3b* in the boundaries contribute to the generation of actomyosin cables that are instrumental in speeding the cell sorting into the different rhombomeres. Shokingly, those cables and the fast sorting mechanism seem to be a specific adaptation of *Ostariophysi* to their rapid embryonic development. Importantly, the acquisition of the new expression domain by *rac3b* was likely caused by the inclusion in the *rac3b/rfng* TAD of a weak enhancer located in the proximity of the sgca promoter, that is conserved outside *Ostariophysi*. This weak enhancer seeded the appearance of additional redundant enhancers that eventually contributed to the strong expression of *rac3b* in the hindbrain boundaries in this group of fishes. The second example is the stepwise origin of the chromatin architecture in two TADs found around the HoxA and HoxD clusters in vertebrates, that is one of the main results of this thesis and will be further discussed in the next section (Acemel et al. 2016).

5.4

A stepwise elaboration of the vertebrate Hox architecture

The collinear expression of at least the HoxA and the HoxD cluster of genes in the limb buds of mammals is a necessary condition in the patterning of the arms, forearms, hands and digits. Such collinearity must be split in two partially overlapping waves: an early-proximal involving anterior and middle Hox genes and a late-distal involving middle and posterior Hox genes (reviewed in Lonfat and Duboule 2015). In order to achieve this split expression, HoxA and HoxD clusters need to be precisely located in the middle of two TADs: an anterior one with early-proximal enhancers and a posterior with late-distal (Montavon et al. 2011). This position allow intermediate Hox genes to be activated both by early-proximal and late-distal enhancers upon a slight change in the 3D architecture (Andrey et al. 2013). The fact that both HoxA and HoxD (and arguably also the HoxB) ohnolog loci display an equivalent configuration in two TADs with the Hox genes in the middle already suggests an ancient origin of this regulatory mechanism that predate the origin of vertebrates and the two rounds of WGDs. Indeed, in teleost fishes hoxD and hoxA clusters also share this peculiar architecture (Woltering et al. 2014) and recently it has also been demonstrated that the two waves of separated Hox expression are also needed for the patterning of fins (Nakamura et al. 2016). Furthermore, in this work we reconstructed the synteny around the four vertebrate Hox clusters and found that it was well preserved among them both anterior and posteriorly. This further indicates that the Hox cluster of the preduplicative ancestor of vertebrates was also in the middle of two TADs and this situation was probably relaxed afterwards in the HoxC cluster. Then the question remained: when and how did this rather complex chromatin organization appear? Is it restricted to vertebrates with paired appendages? In this work we have shed some light by

exploring the architecture of the single Hox locus of the cephalochordate amphioxus, which is informative in order to reconstruct the situation in the last common ancestor of chordates. In contrast to the situation in vertebrates, in amphioxus all Hox genes are embedded within the same TAD that includes the Hox cluster and an anterior flanking region. Remarkably, this anterior flanking region is equivalent in terms of synteny and presumably homologous to the anterior TAD of vertebrates, suggesting that the bipartite architecture found in vertebrates arose in several steps.

Then, with the current data we support a model in which in the bilaterian ancestor none of the two flanking regions was neither structurally nor functionally wired to the regulation of Hox genes (Figure 5.4). This is supported by the fact that there is no apparent microsyntenic constraints indicating a connection between Hox genes and the anterior nor the posterior flanking regions that predate the origin of chordates. Furthermore, we have found that the 3D configuration around the Hox locus of the arthropod *Strigamia maritima* display little contacts with either of the flanking regions. It is plausible and even probable that Hox genes of several extant non chordate lineages interact with enhancers of either their anterior or posterior neighboring regions. However, it is most likely that this was not the configuration in the bilaterian ancestor and those contacts appeared later on independently from the acquisition of the anterior and posterior TADs in the vertebrate lineage.

Later, before the split of the different lineages of chordates, a genomic rearrangement connected the Hox cluster with the genomic region corresponding to the anterior TAD of vertebrates. This connection was presumably functional already in the last common ancestor of chordates, since this anterior region in amphioxus contain enhancers compatible with the Hox expression in the CNS (Pascual-Anaya et al. 2012). Whether this rearrangement had any adaptive value in the first place is difficult to guess. This newly wired anterior region could have harbored enhancers that were readily hijacked by the Hox cluster or could have been populated with Hox-related distal enhancers afterwards. Interestingly, some of the amphioxus enhancers of this region drive expression in vertebrate specific cell populations when introduced in zebrafish, for instance in the neural crest. That might indicate that there are important similarities in the GRN circuitry between the vertebrate neural crest and some populations of the amphioxus CNS, hypothesis that has already been proposed (Manzanares et al. 2000) although it needs to be explored further.

Eventually, another genomic rearrangement brought the genomic region corresponding to the posterior TAD before the two WGDs. Indeed, that region can be traced by synteny in amphioxus, but disconnected from the amphioxus Hox cluster. Additional elaboration occurred, distal limb enhancers appeared, and Hox genes transitioned from being embedded within a single TAD to be placed precisely at the boundary that separates the two different genomic regions: the anterior TAD already wired in the last common ancestor of chordates and the posterior TAD that was linked later on. There are several plausible scenarios but we envision one in which first the anterior and the posterior flanking regions were included within a big single TAD that was bisected in two later on. Nevertheless, an alternative explanation that cannot be totally ruled out is that the last common ancestor of chordates already displayed the configuration in two TADs and the posterior region was secondarily disconnected in amphioxus. However, this seems more unlikely. The Hox cluster of amphioxus is not compact, it spans more than 400 kb. This seems to be the case in most organisms that keep Hox genes in a single cluster, with the exception of vertebrates (Duboule 2007). Meanwhile, in vertebrates the intergenic spacing between Hox genes is drastically smaller so that they 'fit' in the TAD boundary vicinity. A scenario in which the addition of the second TAD was accompanied by the compaction of the cluster to benefit from the two regulatory inputs

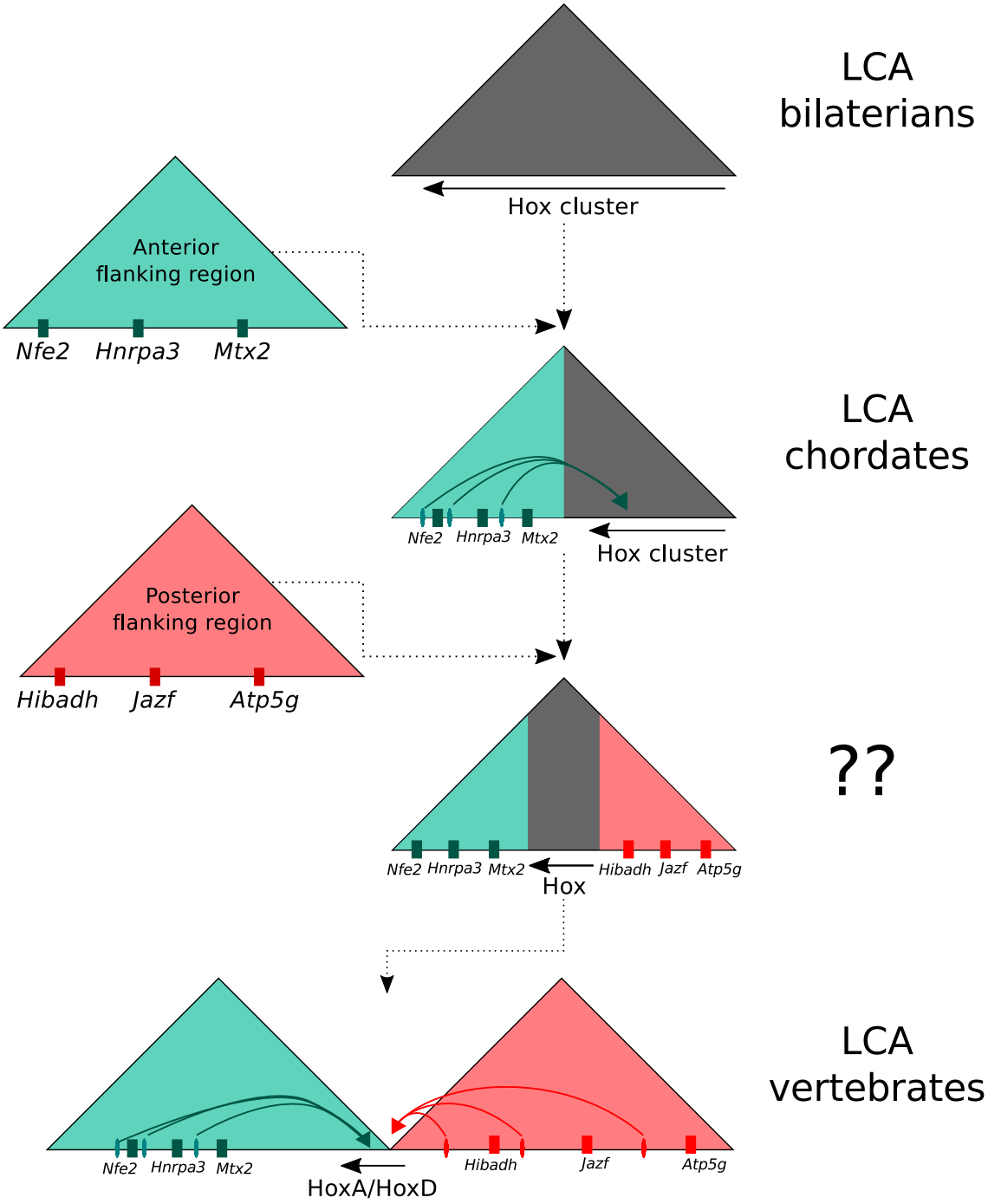


Figure 5.4: A plausible path for the evolution of the vertebrate HoxD chromatin configuration in two TADs. In the bilaterian ancestor no neighboring flanking region was wired to the regulation of Hox genes. Then, the homologous of the vertebrate anterior flanking region got wired and fixed through the appearance of distal enhancers in the LCA of chordates. Later on, before the WGDs, the homologous to the vertebrate posterior neighboring region was wired and the original TAD split in two.

seems more plausible.

Unfortunately, trying to infer how those events happened is challenging due to the scarcity of informative species diverging from the vertebrate lineage in between the last common ancestor of chordates and the last common ancestor of jawed vertebrates. The genomes of tunicates, which are

phylogenetically closer to vertebrates than amphioxus, are too divergent to be informative with the Hox gene cluster scattered in different loci (Ikuta et al. 2004). In contrast, an in depth study of the chromatin folding around the Hox loci in ciclostomes (lampreys and hagfishes), that lack paired appendages, is in need. However, there are some difficulties. Ciclostomes seem to have undergone three rounds of WGDs, but it is still unclear whether it is one or two of them that are shared with jawed vertebrates (Pascual-Anaya et al. 2018). Due to the WGDs, six Hox clusters are found both in lampreys and hagfishes, but the orthology between those clusters and the four clusters observed in most vertebrates is unclear. Indeed, the orthologies between the six clusters of lampreys and the six of hagfishes is not resolved either. In addition, more contiguous genome assemblies are also in need to explore the synteny of neighboring regions and perform informative chromatin conformation capture experiments. This will allow to check for the bipartite configuration in two TADs in ciclostomes. If that was the case, since the genetic program of paired and unpaired fins have been shown to be largely equivalent (Letelier et al. 2018a), it would be interesting to test whether this particular architecture is required for the proper expression of Hox genes in the dorsal fin in the ciclostome taxa.

5.5

Impact of whole genome duplications in the evolution of Regulatory Landscapes

For many years, the paradigm that developmental processes evolve mainly through changes in the transcriptional regulation of otherwise conserved genes has heavily influenced the research in evo-devo (Carroll 2008). Indeed, this work also aims to find regulatory changes that were important in the origin of the morphological novelties of the vertebrate body plan, trying to incorporate the role of changes in the chromatin architecture. However, it is worth noting that regulatory changes are not the only mechanism driving developmental changes. Other mechanisms that have been recently vindicated include the asymmetric evolution of paralog genes after duplications (Holland et al. 2017), which takes advantage of the redundancy that is generated.

Several examples of neofunctionalization of genes after tandem duplications have been reported. For instance, a group of fast-evolving homeobox TFs derived from tandem duplications of the retinal gene *Crx*, that occurred at the origin of placental mammals, now activate a different set of target genes during the early development of mammals (Maeso et al. 2016). These genes might be important for the determination of mammalian extra-embryonic tissues such as the placenta, which is a novelty of this group. Similarly, *Shx* genes generated from tandem duplications of the Hox gene *zen* at the root of lepidopterans (butterflies and moths) have also evolved new roles in early development (Ferguson et al. 2014). They participate in the determination of the cells becoming serosa, which is an extra-embryonic membrane that wraps the embryo protecting it against dessication. This is related to the fact that lepidopterans adapted to lay eggs directly on leaves instead of lay on damp soils. Finally, we will also refer to a tandem duplication occurring in the water strider genus *Rhagovelia* that have been nicely linked to the appearance of a morphological novelty of this group, a fan shaped structure in the second leg (Santos et al. 2017). This structure allow these swimming insects to colonize environments with fast water streams. Although the

asymmetric evolution of the coding sequence of paralog genes and changes in gene regulation are separated from a theoretical point of view, it is worth noting that in all the three cases regulatory and coding changes came hand in hand. Both *Crx* and *zen* derived genes, apart from their fast coding divergence, became expressed earlier on in extraembryonic tissues. Similarly, the duplicate *geisha* become expressed in the second leg primordia in order to generate the fans of *Rhagovelia*, in contrast to her partner *mother of geisha*.

More drastic events are WGDs, that produce a copy of the entire genome (coding and regulatory sequences included). They have occurred at the root of different important eukaryotic lineages, including vertebrates (Dehal and Boore 2005), and have been traditionally linked to a transient period of enhanced evolvability due to genome redundancy (Ohno 1970). Therefore, they have been also historically related to the appearance of morphological novelties, an increased complexity of the body plans and even evolutionary success. Indeed, there are some suggestive examples like the radiation of vertebrates (Marlétaz et al. 2018), eudicot plants (preceded by a whole genome triplication, Ren et al. 2018) and arachnopulmonates (including spiders and scorpions, that adapted to terrestrial environments, Schwager et al. 2017). However, there are also a number of examples where this relationship is less clear. For instance, the radiation of the two most populated teleost taxa (*Ostariphysi* and *Percomorpha*) is not coordinated with the third WGD of the teleost lineage and did not entailed important body plan modifications (Glasauer and Neuhauss 2014). Similarly, the WGDs of horseshoe crabs (Kenny et al. 2016) and rotifers (Flot et al. 2013) did not seem to be accompanied by a huge morphological impact. Nevertheless, it is important to note that the events happening immediately after WGDs are key in order to determine their final influence and it is conceivable that they might be variable.

The newly generated paralog genes can follow different fates, the most important being either to be maintained or to be lost. Indeed, gene losses are frequent upon WGDs, although rather interestingly developmental gene paralogs were retained in several copies in higher proportions than the rest of genes both in vertebrates (Marlétaz et al. 2018) and arachnopulmonates (Schwager et al. 2017). If more than one paralog is retained, there are several concepts that help to classify the behavior of such copies after the WGD event: (I) redundancy imply that all paralogs perform the same functions and share all their expression domains among themselves and with the ancestral copy, (II) sub-functionalization means that they distribute the "duties" (expression domains and functions) of the ancestral copy, in the (III) specialization scenario one of the genes retain all the ancestral functions and domains while the others specialize in some of them and finally, in the (IV) neofunctionalization, one of the paralogs again retains the ancestral activity while the others acquire new functions and expression domains like the case of *geisha* in *Rhagovelia*. Until very recently, exploring the frequency of each of the scenarios after WGDs genome wide was challenging and most studies focused mostly in the mutation rates of the coding sequences of paralog genes (Hu et al. 2016).

Notwithstanding, a recent seminal study has shed light on this question in the particular case of the WGDs of vertebrates, with a special focus on the gene regulation end. Taking advantage of RNAseq experiments performed in a comprehensive collection of adult tissues and developmental stages of both amphioxus and zebrafish, it could be established that specialization was the most frequent scenario after WGDs at the origin of vertebrates (Marlétaz et al. 2018). Rather shockingly most specialized genes, which are the ones that retain less expression domains, tended to be linked to a higher number of regulatory elements. This poses a new and perhaps counter-intuitive concept: an increased regulatory complexity is often required in order to achieve specific and restricted

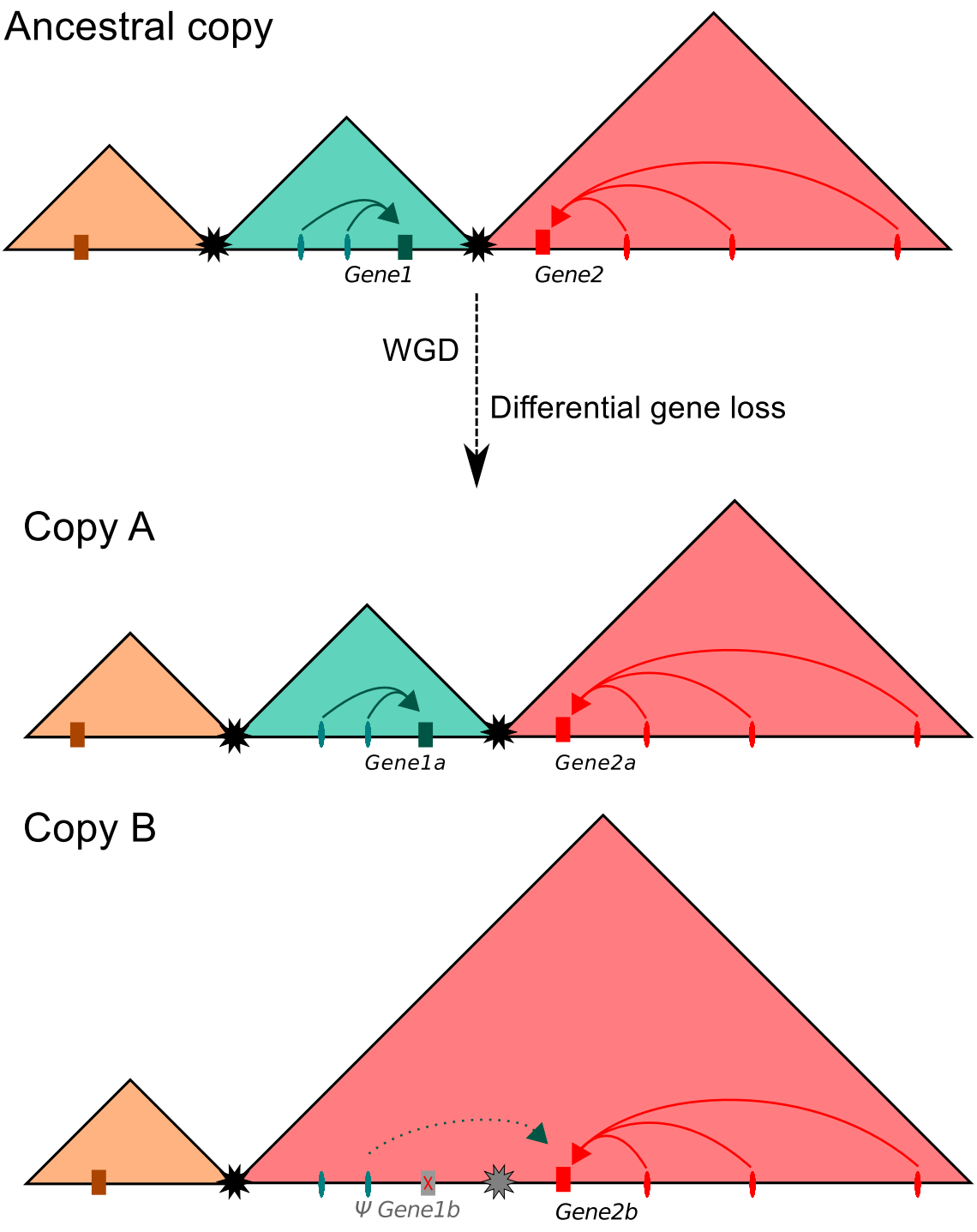


Figure 5.5: Mechanism of relaxation of a TAD boundary following a WGD event. After the WGD, the loss of the *Gene1b* paralog might reduce the pleiotropic effects of losing the TAD boundary separating the blue and the red TAD, therefore facilitating the expansion of the *Gene2b* RL.

expression domains. Such strongly specialized genes with many associated enhancers were mainly expressed in the brain, which might indicate that some of the novelties found in vertebrate neural tube development were favored by the WGDs.

In the same study, using ATAC-seq experiments in zebrafish and amphioxus embryos to identify enhancers, an interesting pattern of paralog specific increase in regulatory complexity was observed.

This pattern is likely in concordance with the scenario in which specialization was the predominant fate of the different paralogy groups in the origin of vertebrates. Indeed, when a given paralogy group is retained as a singleton in vertebrates, the number of associated enhancers tend to be slightly higher in zebrafish than in amphioxus. However, if more than one copy is retained, one of the vertebrate paralogs tend to display the same number of regulatory elements than amphioxus (often the one retaining the ancestral functions) while the other copies gain an important number of enhancers (often the specialized ones). However, it is worth noting that the enhancer-promoter assignation in this study was not based on chromatin architecture, but was performed using the *GREAT* bioinformatic tool (McLean et al. 2010). This tool assign enhancers to genes based on linear proximity, and the fact that intergenic distances in zebrafish (and in vertebrates in general) are larger than in amphioxus could be biasing the analysis. However, in this thesis we repeated this analysis taking chromatin architecture into account by using the RL sizes calculated from the H3K4me3 HiChIP experiments, and found an strikingly similar pattern. Furthermore, the very same pattern could be also observed if we used the RL prediction to assign the ATAC-seq peaks to their target promoters. Finally, those paralog genes that experienced higher growths in their RLs were predominantly associated to both neural functions and expression domains, in concordance with the previous findings. Apart from giving additional support to the findings of the aforementioned study, it is worth noting that this work is the first study that directly address the question of how chromatin architecture evolve in response to a WGD.

However, the question of whether WGDs influenced or not the growth of the RLs of specialized genes, and if so how did it happen, is still open. Notwithstanding, there is a number of plausible mechanisms relating both things, and most likely the increase of intergenic distances that derived from the pervasive differential gene losses was one of them. In this work we have found that the two gene deserts observed around the HoxD cluster in vertebrates, for instance, are most likely the result of erasing the coding sequences of syntenic paralogs that are retained in other clusters. In addition, it is tempting to speculate that upon the loss of the genes of a given TAD that respond to distal regulation, structural variants altering the TAD boundaries and linking the 'orphan' enhancers to a new promoter might be less pleiotropic (Figure 5.5). Such variants might still have potential deleterious effects resulting from the gain of function phenotype of the newly wired gene, but the loss of function effects that may derive from disconnecting the only copy of a particular gene from its enhancers or by the introduction of a competitor promoter in the TAD might be buffered away. Indeed, even without considering the hijacking of preexisting enhancers, the erosion of redundant TADs (genes and some of the enhancers included) followed by structural variants connecting them to other TADs might be an overlooked mechanism for the growth and evolution of RLs upon a WGD. Also, actively transcribed genes are enriched at TAD boundaries and active transcription seems to be an evolutionary conserved mechanism driving chromatin insulation (Rowley et al. 2017). Then, after WGDs, the elimination of redundant highly expressed genes could have led to the relaxation of some boundaries.

5.6

A different compartment for the Polycomb mediated long range contacts in vertebrates

So far we have explored regulatory changes between vertebrates and cephalochordates under the assumption that regulatory mechanisms were largely equivalent, and that only the way they were deployed was susceptible to change. This is generally true, for instance in both species open chromatin regions and the H3K27ac epigenetic mark are both predictive of active promoters and enhancers while H3K4me3 punctuate exclusively active promoters. Furthermore, in this thesis we have shown how chromatin architecture features like TADs and A/B compartments are also present in cephalochordates, and operate in a highly similar manner.

However, the H3K27me3 epigenetic mark, that is indicative of facultative repression through the Polycomb group of proteins, seems to be distributed differently in vertebrates. In contrast to the situation in other metazoans, according to our H3K27me3 HiChIP experiments and to previously published ChIP-seq data, the H3K27me3 signal is sharply deposited around promoters in 24hpf zebrafish embryos. In contrast, Polycomb domains both in amphioxus and in *Drosophila* are much wider, as it has been shown also both with HiChIP and regular ChIP-seqs. This is true for both species even though whole 15hpf embryos were used for the amphioxus experiments and the embryonic Kc167 cell line was used for the *Drosophila* ones. This rule out that the wider distribution of the H3K27me3 signal outside vertebrates is due to the cellular heterogeneity of the material used, together with the fact that the sharp H3K27me3 peaks observed in zebrafish also come from experiments performed in whole embryos. Rather, the striking differences in the methylation frequency of CpG islands observed in vertebrates (Feng et al. 2010) when compared to other groups might explain such differences, since H3K27me3 and 5mC methlyation display complementary patterns. While most vertebrate CpG islands are methylated (Bogdanović et al. 2016), DNA methylation seems to be either absent or residual in *Drosophila* (Dunwell and Pfeifer 2014) and concentrated in the gene bodies of actively transcribed genes in amphioxus (Marlétaz et al. 2018). Then, the expansion of DNA methylation in vertebrates might be sufficient or at least could partially explain the reduction in size of the Polycomb domains in this lineage.

In addition, our H3K27me3 HiChIP experiments in zebrafish embryos revealed an increase in the long range interactions of the promoters associated with Polycomb when compared with the same promoters when they are associated with H3K4me3. Such long range interactions seem to connect Polycomb domains belonging to different TADs among themselves and interestingly some of those interactions are evolutionarily conserved. For instance, the Polycomb domains found at the promoters of the zebrafish hoxD genes interact with the one located around the *sp9* gene and with the ones around the promoters of the *dlx1a* and *dlx2a* genes. Such contacts were also described for the orthologous genes in mice using 4C-seq in embryonic dissections of forebrain cells, where all those genes are silent (Vieux-Rochas et al. 2015). However, whether such long range interactions that overflow the limits of TAD boundaries play any role in gene regulation remains to be elucidated with functional assays. Such TAD crossing contacts are also observable in the H3K27me3 HiChIP experiments performed in *Drosophila*, but the sizes of the Polycomb domains involved are larger.

Since A/B compartments are the next feature in the hierarchy of chromatin organization we

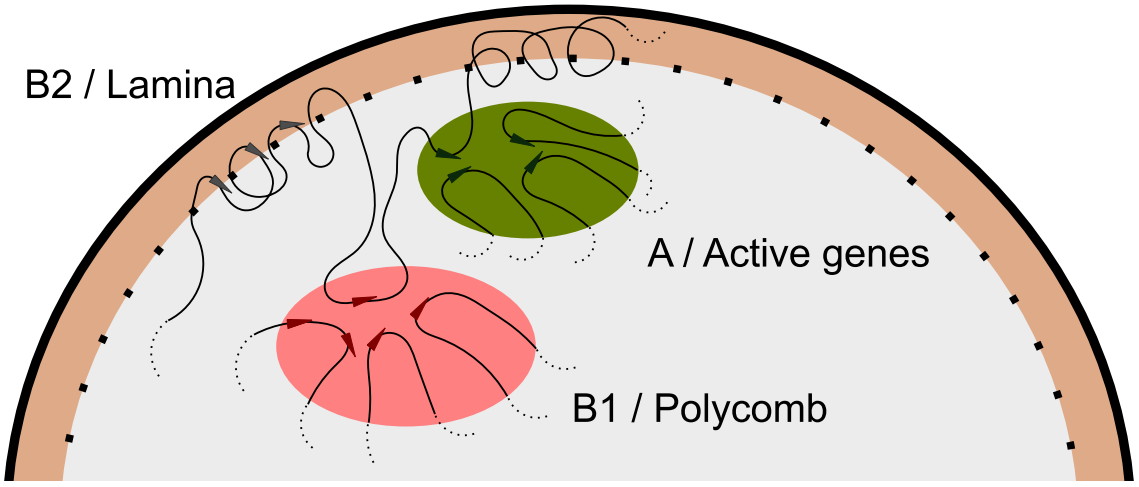


Figure 5.6: Cartoon representing three of the sub-compartments proposed for mammalian cells in Rao et al. 2014. In the B1 compartment, Polycomb repressed or poised developmental genes cluster together. In the A, active genes are reunited to be transcribed. In the B2 compartment, more stable repressed genes and repetitive elements are associated to the nuclear lamina and kept silent.

wondered how Polycomb interactions related to compartments. Intriguingly, while in *Drosophila* it has been well described that Polycomb domains are associated with the majority of the regions belonging to the B compartment, in zebrafish this is unclear. Remarkably, Polycomb domains are located near boundaries in between the two compartments, but always towards the A side. In principle this is at odds with the classic notion in which the chromatin that is not transcribed belong to the B compartment. However, the precise location in the vicinity of A to B switches could be supporting the existence of a third and independent compartment tied to facultative repression through Polycomb. Indeed, an independent compartment name for H3K27me3-rich areas has already been proposed using high resolution HiC experiments in human cell lines, termed the B1 compartment (Rao et al. 2014, Figure 5.6). Perhaps this B1 compartment also exists in zebrafish cells and tend to interact more with regions of the A compartment than with the lamina associated repressed chromatin of the so called B2 compartment also described in human cell lines. Or perhaps, given that we are in the context of a heterogeneous embryo, regions that belong to the B1 compartment in some cells might switch to the A compartment in others with higher frequency than regions associated to lamina that belong to a more stable B2 compartment. Whether this subdivision of the B compartment is exclusive of vertebrates and linked to the changes in the patterns of DNA methylation and H3K27me3 occupancy needs to be fully addressed. Unfortunately, the contiguity of the amphioxus assembly is still insufficient to explore such long range interactions in detail and predict the situation in the chordate ancestor. In addition, the case of *Drosophila* where H3K27me3 marks most of the B compartment might be a derived situation that does not represent the ancestral state of protostomes and further chromosome conformation capture studies in other protostome models are needed. However, our data suggest that such division is conserved at least across gnathostomes, from mammals to teleost fishes. In addition, we have also provided the first proof that A and B compartments also operate in the cephalochordate lineage.

Chapter 6

Conclussions/Conclusiones

1. Changes in the folding of the genome were required for the evolution of regulatory novelties of the vertebrate lineage, as shown with comparative 4C-seq and HiChIP experiments.

2. The chromatin organization around the vertebrate HoxD locus, with Hox genes located precisely at the boundary between two TADs, originated in a stepwise manner. The anterior TAD was functionally wired to the regulation of Hox genes already in the LCA of chordates. Meanwhile, the posterior TAD was connected later on but before the first round of WGD that happened in the vertebrate ancestor.

3. 393 cases of vertebrate specific genomic rearrangements that altered the integrity of TADs and therefore potentially generated changes in gene regulation were identified using comparative H3K4me3 HiChIP experiments.

4. Ohnolog Regulatory Landscapes (RLs) that originated after the two rounds of WGDs evolved differently depending of the pattern of paralog retention. RLs retained in more than one copy tend to follow a pattern in which one of the copies retain the ancestral size and the others grow both in size and in number of CREs associated.

5. The long range contacts observed between inactive developmental promoters in vertebrates, likely mediated by Polycomb, seem to be a novelty of this lineage. This might be related to a vertebrate specific subdivision of the B compartment.

1. Para la aparición de algunas de las novedades evolutivas de vertebrados se necesitaron cambios en la arquitectura tridimensional del genoma.

2. La organización tridimensional de la cromatina alrededor de los genes HoxD en vertebrados, con los genes Hox localizados en el borde entre dos *TADs*, apareció de forma secuencial. El *TAD* anterior estaba ya ligado a la regulación de los genes Hox en el ancestro de cordados mientras que el posterior quedó conectado después, pero antes de la primera ronda de duplicación de genoma completo ocurrida en el ancestro de vertebrados.

3. Usando experimentos de HiChIP se identificaron 393 casos de rearreglos cromosómicos que potencialmente pudieron alterar la integridad de los TADs en el linaje de los vertebrados y por tanto pudieron causar cambios de expresión génica.

4. Los paisajes reguladores de genes parálagos originados tras las dos rondas de duplicación de genoma completo evolucionaron de forma diferente, dependiendo del número de parálogos conservados. Los paisajes reguladores retenidos en más de una copia siguen un patrón en el que uno de ellos retiene el tamaño ancestral y el resto crecen tanto en tamaño como en número de elementos cis-reguladores.

5. Los contactos de larga distancia que se observan entre promotores de genes de desarrollo inactivos en vertebrados, probablemente mediados por proteínas de la familia Polycomb, parecen ser una novedad de este linaje. Esto podría tener relación con una subdivisión del compartimento B (inactivo) específica de vertebrados.

Bibliography

Abitua, P. B. et al. (2015). "The pre-vertebrate origins of neurogenic placodes". In: *Nature* 524.7566, pp. 462–465.

Acemel, R. D. et al. (2016). "A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation." In: *Nature genetics* 48.3, pp. 336–341.

Aires, R. et al. (2016). "Oct4 Is a Key Regulator of Vertebrate Trunk Length Diversity". In: *Developmental Cell* 38.3, pp. 262–274.

Akalin, A. et al. (2009). "Transcriptional features of genomic regulatory blocks." In: *Genome biology* 10.4, R38.

Albuixech-Crespo, B. et al. (2017). "Molecular regionalization of the developing amphioxus neural tube challenges major partitions of the vertebrate brain." In: *PLoS biology* 15.4, e2001573.

Alexander, J. et al. (1999). "casanova Plays an early and essential role in endoderm formation in zebrafish". In: *Developmental Biology* 215.2, pp. 343–357.

Alexander, T., C. Nolte, and R. Krumlauf (2009). "Hox Genes and Segmentation of the Hindbrain and Axial Skeleton ". In: *Annual Review of Cell and Developmental Biology* 25.1, pp. 431–456.

Altenhoff, A. M. et al. (2018). "The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces". In: *Nucleic Acids Research* 46.D1, pp. D477–D485.

Amemiya, C. T. et al. (2013). "The African coelacanth genome provides insights into tetrapod evolution". In: *Nature* 496.7445, pp. 311–316.

Amores, A. (1998). "Zebrafish hox Clusters and Vertebrate Genome Evolution". In: *Science* 282.5394, pp. 1711–1714.

Andrey, G. et al. (2013). "A switch between topological domains underlies HoxD genes collinearity in mouse limbs." In: *Science* 340.6137, p. 1234167.

Aulehla, A. and O. Pourquié (2010). "Signaling gradients during paraxial mesoderm development." In: *Cold Spring Harbor perspectives in biology* 2.2, a000869.

Avery, O. T., C. M. MacLeod, and M. McCarty (1944). "Studies on the chemical nature of the substance inducing transformation of pneumococcal types". In: *Journal of Experimental Medicine* 79.2, pp. 137–159.

Ay, F. et al. (2014). "Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression". In: *Genome Research* 24.6, pp. 974–988.

Bailey, S. D. et al. (2015). "ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters". In: *Nature Communications* 2.

Barbosa-Morais, N. L. et al. (2012). "The evolutionary landscape of alternative splicing in vertebrate species." In: *Science* 338.6114, pp. 1587–1593.

Bassham, S. and J. Postlethwait (2000). "Brachyury (T) expression in embryos of a larvacean urochordate, Oikopleura dioica, and the ancestral role of T". In: *Developmental Biology* 220.2, pp. 322–332.

Beaster-Jones, L. et al. (2008). "Expression of somite segmentation genes in amphioxus: A clock without a wavefront?" In: *Development Genes and Evolution* 218.11-12, pp. 599–611.

Beccari, L. et al. (2016). "A role for HOX13 proteins in the regulatory switch between TADs at the HoxD locus." In: *Genes & development* 30.10, pp. 1172–86.

Becker, J. S., D. Nicetto, and K. S. Zaret (2016). "H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes". In: *Trends in Genetics* 32.1, pp. 29–41.

Bessa, J. et al. (2002). "Combinatorial control of Drosophila eye development by eyeless, homothorax, and teashirt." In: *Genes {&} development* 16.18, pp. 2415–2427.

Bintu, B. et al. (2018). "Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells". In: *Science* 362.6413, eaau1783.

Bogdanović, O. and R. Lister (2017). "DNA methylation and the preservation of cell identity". In: *Current Opinion in Genetics and Development* 46, pp. 9–14.

Bogdanovic, O. et al. (2012). "Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis". In: *Genome Research* 22.10, pp. 2043–2053.

Bogdanović, O. et al. (2016). "Active DNA demethylation at enhancers during the vertebrate phylotypic period". In: *Nature Genetics* 48.4, pp. 417–426.

Braasch, I. et al. (2016). "The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons". In: *Nature Genetics* 48.4, pp. 427–437.

Brena, C. et al. (2006). "Expression of trunk Hox genes in the centipede Strigamia maritima: sense and anti-sense transcripts." In: *Evolution & development* 8.3, pp. 252–65.

Brenner, S., F. Jacob, and M. Meselson (1961). "An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis". In: *Nature* 190.4776, pp. 576–581.

Brent, A. E. and C. J. Tabin (2002). "Developmental regulation of somite derivatives: Muscle, cartilage and tendon". In: *Current Opinion in Genetics and Development* 12.5, pp. 548–557.

Brookes, E. and A. Pombo (2009). "Modifications of RNA polymerase II are pivotal in regulating gene expression states". In: *EMBO Reports* 10.11, pp. 1213–1219.

Buenrostro, J. D. et al. (2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." In: *Nature methods* 10.12, pp. 1213–8.

Buenrostro, J. D. et al. (2015). "Single-cell chromatin accessibility reveals principles of regulatory variation." In: *Nature* 523.7561, pp. 486–490.

Burke, T. W. and J. T. Kadonaga (1996). "Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters." In: *Genes {&} development* 10.6, pp. 711–724.

Buttler, A. et al. (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74.

Calle-Mustienes, E. de la et al. (2005). "A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts." In: *Genome research* 15.8, pp. 1061–1072.

Canela, A. et al. (2017). "Genome Organization Drives Chromosome Fragility." In: *Cell* 170.3, 507–521.e18.

Cannavò, E. et al. (2016). "Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks". In: *Current Biology* 26.1, pp. 38–51.

Carapuço, M. et al. (2005). "Hox genes specify vertebral types in the presomitic mesoderm". In: *Genes and Development* 19.18, pp. 2116–2121.

Carninci, P. et al. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." In: *Nature genetics* 38.6, pp. 626–635.

Carroll, S. B. (2008). "Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution". In: *Cell* 134.1, pp. 25–36.

Chambeyron, S. (2005). "Nuclear re-organisation of the Hoxb complex during mouse embryonic development". In: *Development* 132.9, pp. 2215–2223.

Chambeyron, S. and W. A. Bickmore (2004). "Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription". In: *Genes & Development* 18.10, pp. 1119–1130.

Choi, S. H. et al. (2016). "DUX4 recruits p300/CBP through its C-terminus and induces global H3K27 acetylation changes". In: *Nucleic Acids Research* 44.11, pp. 5161–5173.

Clark, E. G. et al. (2019). "The function of the ophiuroid nerve ring: how a decentralized nervous system controls coordinated locomotion." In: *The Journal of experimental biology* 222.Pt 2, jeb192104.

Clark-Hachtel, C. M. and Y. Tomoyasu (2016). "Exploring the origin of insect wings from an evo-devo perspective." In: *Current opinion in insect science* 13.2, pp. 77–85.

Crane, E. et al. (2015). "Condensin-driven remodelling of X chromosome topology during dosage compensation". In: *Nature* 523.7559, pp. 240–244.

Cremer, T. and M. Cremer (2010). "Chromosome territories." In: *Cold Spring Harbor perspectives in biology* 2.3, pp. 1–23.

Creyghton, M. P. et al. (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proceedings of the National Academy of Sciences* 107.50, pp. 21931–21936.

Cusanovich, D. A. et al. (2018). "The cis-regulatory dynamics of embryonic development at single-cell resolution". In: *Nature* 555.7697, pp. 538–542.

Darwin, C. (1861). *On the Origin of Species by means of natural selection*. New York: D. Appleton & Company.

Davidson, E. H. and D. H. Erwin (2006). "Gene regulatory networks and the evolution of animal body plans." In: *Science* 311.5762, pp. 796–800.

Dehal, P. and J. L. Boore (2005). "Two rounds of whole genome duplication in the ancestral vertebrate." In: *PLoS biology* 3.10, e314.

Dekker, J. (2002). "Capturing Chromosome Conformation". In: *Science* 295.5558, pp. 1306–1311.

Delsuc, F. et al. (2006). "Tunicates and not cephalochordates are the closest living relatives of vertebrates". In: *Nature* 439.7079, pp. 965–968.

Deng, W. et al. (2012). "Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor". In: *Cell* 149.6, pp. 1233–1244.

Denker, A. and W. De Laat (2016). "The second decade of 3C technologies: Detailed insights into nuclear organization". In: *Genes and Development* 30.12, pp. 1357–1382.

Deutsch, J. S. and E. Mouchel-Vielh (2003). "Hox genes and the crustacean body plan". In: *BioEssays* 25.9, pp. 878–887.

Diogo, R. et al. (2015). "A new heart for a new head in vertebrate cardiopharyngeal evolution". In: *Nature* 520.7548, pp. 466–473.

Dixon, J. R. et al. (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions". In: *Nature* 485.7398, pp. 376–380.

Dixon, J. R. et al. (2015). "Chromatin architecture reorganization during stem cell differentiation". In: *Nature* 518.7539, pp. 331–336.

Domazet-Lošo, T. and D. Tautz (2010). "A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns." In: *Nature* 468.7325, pp. 815–818.

Dominguez, P., A. G. Jacobson, and R. P. Jefferies (2002). "Paired gill slits in a fossil with a calcite skeleton". In: *Nature* 417.6891, pp. 841–844.

Dominguez-Cejudo, M. A. and F Casares (2015). "Anteroposterior patterning of Drosophila ocelli requires an anti-repressor mechanism within the hh pathway mediated by the Six3 gene Optix". In: *Development* 142.16, pp. 2801–2809.

Dostie, J. et al. (2006). "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements." In: *Genome research* 16.10, pp. 1299–309.

Duan, Z. et al. (2010). "A Three-Dimensional Model of the Yeast Genome". In: *Nature* 465.7296, pp. 363–367.

Duboule, D. and P. Dollé (1989). "The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes." In: *The EMBO Journal* 8.5, pp. 1497–1505.

Duboule, D. (1994). "Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony." In: *Development* 42, pp. 135–42.

— (2007). "The rise and fall of Hox gene clusters." In: *Development* 134.14, pp. 2549–60.

Dudchenko, O. et al. (2017). "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds". In: *Science* 356.6333, pp. 92–95.

Dunwell, T. L. and G. P. Pfeifer (2014). "Drosophila genomic methylation: new evidence and new questions". In: *Epigenomics* 6.5, pp. 459–461.

Durand, N. C. et al. (2016). "Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom". In: *Cell Systems* 3.1, pp. 99–101.

Elosegui-Artola, A. et al. (2017). "Force Triggers YAP Nuclear Entry by Regulating Transport across Nuclear Pores". In: *Cell* 171.6, 1397–1410.e14.

Erwin, D. H. et al. (2012). "The Cambrian Conundrum : Early Success in the Early History of Animals". In: 1091.2011, pp. 1091–1098.

Escriva, H. (2018). "My Favorite Animal, Amphioxus: Unparalleled for Studying Early Vertebrate Evolution". In: *BioEssays* 40.12, pp. 1–8.

Fabre, P. J. et al. (2017). "Large scale genomic reorganization of topological domains at the HoxD locus". In: *Genome Biology* 18.1, pp. 1–15.

Farrell, J. A. et al. (2018). "Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis". In: *Science* 360.6392, eaar3131.

Feng, S et al. (2010). "Conservation and divergence of methylation patterning in plants and animals". In: *Proceedings of the National Academy of Sciences* 107.19, pp. 8689–8694.

Ferguson, L. et al. (2014). "Ancient expansion of the hox cluster in lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation." In: *PLoS genetics* 10.10, e1004698.

Flot, J.-F. et al. (2013). "Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga". In: *Nature* 500.7463, pp. 453–457.

Franke, M. et al. (2016). "Formation of new chromatin domains determines pathogenicity of genomic duplications." In: *Nature* 538.7624, pp. 265–269.

Frazer, K. A. et al. (2004). "VISTA: Computational tools for comparative genomics". In: *Nucleic Acids Research* 32.WEB SERVER ISS. Pp. 273–279.

Freitas, R., J. L. Gómez-Skarmeta, and P. N. Rodrigues (2014). "New frontiers in the evolution of fin development". In: *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 322.7, pp. 540–552.

Freitas, R., G. J. Zhang, and M. J. Cohn (2006). "Evidence that mechanisms of fin development evolved in the midline of early vertebrates". In: *Nature* 442.7106, pp. 1033–1037.

Freitas, R. et al. (2012). "Hoxd13 Contribution to the Evolution of Vertebrate Appendages". In: *Developmental Cell* 23.6, pp. 1219–1229.

Fritsch, M. et al. (2015). "Unexpected co-linearity of Hox gene expression in an aculiferan mollusk". In: *BMC Evolutionary Biology* 15.1, pp. 1–17.

Fröbius, A. C., D. Q. Matus, and E. C. Seaver (2008). "Genomic organization and expression demonstrate spatial and temporal Hox gene colinearity in the lophotrochozoan Capitella sp. I". In: *PLoS ONE* 3.12.

Fuglerud, B. M. et al. (2018). "The pioneer factor activity of c-Myb involves recruitment of p300 and induction of histone acetylation followed by acetylation-induced chromatin dissociation". In: *Epigenetics and Chromatin* 11.1, pp. 1–15.

Fullwood, M. J. et al. (2009). "An oestrogen-receptor-alpha-bound human chromatin interactome." In: *Nature* 462.7269, pp. 58–64.

Fulton, D. L. et al. (2009). "TFCat: the curated catalog of mouse and human transcription factors." In: *Genome biology* 10.3, R29.

Galant, R. and S. B. Carroll (2002). "Evolution of a transcriptional repression domain in an insect Hox protein". In: *Nature* 415.6874, pp. 910–913.

Galis, F. and J. A. J. Metz (2001). "Testing the vulnerability of the phylotypic stage: On modularity and evolutionary conservation". In: *Journal of Experimental Zoology* 291.2, pp. 195–204.

Gee, H. (2018). *Across the Bridge: Understanding the Origin of the Vertebrates*. University of Chicago Press.

Geeven, G. et al. (2018). "peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data." In: *Nucleic acids research* 46.15, e91.

Gehrke, A. R. et al. (2014). "Deep conservation of wrist and digit enhancers in fish". In: *Proceedings of the National Academy of Sciences* 112.3, p. 201420208.

Gendron-Maguire, M. et al. (1993). "Hoxa-2 mutant mice exhibit homeotic transformation of skeletal elements derived from cranial neural crest". In: *Cell* 75.7, pp. 1317–1331.

Gibcus, J. H. et al. (2018). "A pathway for mitotic chromosome formation". In: *Science* 359.6376, eaao6135.

Giraldez, A. J. (2006). "Zebrafish MiR-430 Promotes Deadenylation and Clearance of Maternal mRNAs". In: *Science* 312.5770, pp. 75–79.

Glasauer, S. M. and S. C. Neuhauss (2014). "Whole-genome duplication in teleost fishes and its evolutionary consequences". In: *Molecular Genetics and Genomics* 289.6, pp. 1045–1060.

Gómez-Marín, C. et al. (2015). "Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.24, pp. 7542–7547.

Gompel, N. et al. (2005). "Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila". In: *Nature* 433.7025, pp. 481–487.

González-Aguilera, C. et al. (2014). "Genome-wide analysis links emerin to neuromuscular junction activity in Caenorhabditis elegans". In: *Genome Biology* 15.2, R21.

Greene, N. D. E. et al. (2017). "Neural tube closure: cellular, molecular and biomechanical mechanisms". In: *Development* 144.4, pp. 552–566.

Grossman, A. D., J. W. Erickson, and C. A. Gross (1984). "The htpR gene product of E. coli is a sigma factor for heat-shock promoters". In: *Cell* 38.2, pp. 383–390.

Guerreiro, I. et al. (2016). "Reorganisation of Hoxd regulatory landscapes during the evolution of a snake-like body plan". In: *eLife* 5, pp. 1–23.

Haarhuis, J. H. et al. (2017). "The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension". In: *Cell* 169.4, 693–707.e14.

Haberle, V. and B. Lenhard (2016). "Promoter architectures and developmental gene regulation". In: *Seminars in Cell and Developmental Biology* 57, pp. 11–23.

Halder, G, P Callaerts, and W. Gehring (1995). "Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila". In: *Science* 267.5205, pp. 1788–1792.

Hargan-Calvopina, J. et al. (2016). "Stage-Specific Demethylation in Primordial Germ Cells Safeguards against Precocious Differentiation". In: *Developmental Cell* 39.1, pp. 75–86.

Harmston, N. et al. (2017). "Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation." In: *Nature communications* 8.1, p. 441.

Heger, P., B. Marin, and E. Schierenberg (2009). "Loss of the insulator protein CTCF during nematode evolution." In: *BMC molecular biology* 10.1, p. 84.

Heger, P. et al. (2012). "The chromatin insulator CTCF and the emergence of metazoan diversity." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.43, pp. 17507–12.

Hnisz, D. et al. (2016). "Activation of proto-oncogenes by disruption of chromosome neighborhoods." eng. In: *Science* 351.6280, pp. 1454–1458.

Holland, N. D., L. Z. Holland, and P. W. H. Holland (2015). "Scenarios for the making of vertebrates." In: *Nature* 520.7548, pp. 450–455.

Holland, P. W. H. et al. (2017). "New genes from old: asymmetric divergence of gene duplicates and the evolution of development." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 372.1713, p. 20150480.

Holland, P. W. (2013). "Evolution of homeobox genes". In: *Wiley Interdisciplinary Reviews: Developmental Biology* 2.1, pp. 31–45.

Hoskins, R. A. et al. (2011). "Genome-wide analysis of promoter architecture in Drosophila melanogaster". In: *Genome Research* 21.2, pp. 182–192.

Hou, C. et al. (2012). "Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains". In: *Molecular Cell* 48.3, pp. 471–484.

Howe, D. G. et al. (2012). "ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics". In: *Nucleic Acids Research* 41.D1, pp. D854–D860.

Hsieh, T. H. S. et al. (2015). "Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C". In: *Cell* 162.1, pp. 108–119.

Hu, Y. et al. (2016). "The Atlantic salmon genome provides insights into rediploidization". In: *Nature* 533.7602, pp. 200–205.

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources". In: *Nature Protocols* 4.1, pp. 44–57.

Hudson, C. (2016). "The central nervous system of ascidian larvae". In: *Wiley Interdisciplinary Reviews: Developmental Biology* 5.5, pp. 538–561.

Hueber, S. D. et al. (2010). "Improving Hox Protein Classification across the Major Model Organisms". In: *PLoS ONE* 5.5. Ed. by R. DeSalle, e10820.

Iborra, F. J. et al. (1996). "Active RNA polymerases are localized within discrete transcription "factories' in human nuclei." In: *Journal of cell science* 109.5, pp. 1427–36.

Ikuta, T. et al. (2004). "Ciona intestinalis Hox gene cluster: Its dispersed structure and residual colinear expression in development". In: *Proceedings of the National Academy of Sciences* 101.42, pp. 15118–15123.

Irastorza-Azcarate, I. et al. (2018). "4Cin: A computational pipeline for 3D genome modeling and virtual Hi-C analyses from 4C data". In: *PLoS Computational Biology* 14.3, pp. 1–20.

Irie, N. and S. Kuratani (2014). "The developmental hourglass model: a predictor of the basic body plan?" In: *Development* 141.24, pp. 4649–4655.

Irimia, M. et al. (2010). "Conserved developmental expression of Fezf in chordates and Drosophila and the origin of the Zona Limitans Intrathalamica (ZLI) brain organizer." In: *EvoDevo* 1.1, p. 7.

Irimia, M. et al. (2012). "Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints". In: *Genome Research* 22.12, pp. 2356–2367.

Jacob, F. and J. Monod (1961). "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3.3, pp. 318–356.

Jäger, R. et al. (2015). "Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci". In: *Nature Communications* 6.1, p. 6178.

Jandzik, D. et al. (2015). "Evolution of the new vertebrate head by co-option of an ancient chordate skeletal tissue". In: *Nature* 518.7540, pp. 534–537.

Jeffery, W. R., A. G. Strickler, and Y. Yamamoto (2004). "Migratory neural crest-like cells form body pigmentation in a urochordate embryo". In: *Nature* 431.7009, pp. 696–699.

Johnson, D. S. et al. (2007). "Genome-Wide Mapping of in Vivo Protein-DNA Interactions". In: *Science* 316.5830, pp. 1497–1502.

Jozwik, K. M. et al. (2016). "FOXA1 Directs H3K4 Monomethylation at Enhancers via Recruitment of the Methyltransferase MLL3." In: *Cell reports* 17.10, pp. 2715–2723.

Kaaij, L. J. et al. (2018). "Systemic Loss and Gain of Chromatin Architecture throughout Zebrafish Development". In: *Cell Reports*, pp. 1–10.

Kalinka, A. T. et al. (2010). "Gene expression divergence recapitulates the developmental hourglass model". In: *Nature* 468.7325, pp. 811–814.

Kawakami, K. (2004). "Transgenesis and Gene Trap Methods in Zebrafish by Using the Tol2 Transposable Element". In: *Methods in Cell Biology* 77, pp. 201–222.

Kenny, N. J. et al. (2016). "Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs". In: *Heredity* 116.2, pp. 190–199.

Kerpedjiev, P. et al. (2018). "HiGlass: web-based visual exploration and analysis of genome interaction maps." In: *Genome biology* 19.1, p. 125.

Knight, R. D. et al. (2000). "An amphioxus Krox gene: Insights into vertebrate hindbrain evolution". In: *Development Genes and Evolution* 210.10, pp. 518–521.

Krefting, J., M. A. Andrade-Navarro, and J. Ibn-Salem (2018). "Evolutionary stability of topologically associating domains is associated with conserved gene regulation". In: *BMC Biology* 16.1, p. 87.

Le, T. B. K. et al. (2013). "High-resolution mapping of the spatial organization of a bacterial chromosome." In: *Science* 342.6159, pp. 731–4.

Lee, W. et al. (2007). "A high-resolution atlas of nucleosome occupancy in yeast". In: *Nature Genetics* 39.10, pp. 1235–1244.

Letelier, J. et al. (2018a). "A conserved Shh cis-regulatory module highlights a common developmental origin of unpaired and paired fins". In: *Nature Genetics* 50.4, pp. 504–509.

Letelier, J. et al. (2018b). "Evolutionary emergence of the rac3b/rfng/sgca regulatory cluster refined mechanisms for hindbrain boundaries formation." In: *Proceedings of the National Academy of Sciences of the United States of America* 115.16, E3731–E3740.

Levin, M. et al. (2012). "Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo". In: *Developmental Cell* 22.5, pp. 1101–1108.

Levin, M. et al. (2016). "The mid-developmental transition and the evolution of animal body plans". In: *Nature* 531.7596, pp. 637–641.

Lewis, E. B. (1978). "A gene complex controlling segmentation in Drosophila." In: *Nature* 276.5688, pp. 565–70.

Li, Q. et al. (2011). "Measuring reproducibility of high-throughput experiments". In: *Annals of Applied Statistics* 5.3, pp. 1752–1779.

Lieberman-Aiden, E. et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." In: *Science (New York, N.Y.)* 326.5950, pp. 289–93.

Lin, Y. S. et al. (2009). "Characterization of SoxB2 and SoxC genes in amphioxus (Branchiostoma belcheri): Implications for their evolutionary conservation". In: *Science in China, Series C: Life Sciences* 52.9, pp. 813–822.

Liu, H. et al. (2014). "Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies". In: *Genome Biology and Evolution* 6.3, pp. 510–525.

Lo, P. C. et al. (2002). "Homeotic genes autonomously specify the anteroposterior subdivision of the Drosophila dorsal vessel into aorta and heart". In: *Developmental Biology* 251.2, pp. 307–319.

Local, A. et al. (2018). "Identification of H3K4me1-associated proteins at mammalian enhancers". In: *Nature Genetics* 50.1, pp. 73–82.

Lonfat, N. and D. Duboule (2015). "Structure, function and evolution of topologically associating domains (TADs) at HOX loci." eng. In: *FEBS Letters* 589.20, pp. 2869–2876.

Lowe, C. J. et al. (2003). "Anteroposterior patterning in hemichordates and the origins of the chordate nervous system". In: *Cell* 113.7, pp. 853–865.

Lowe, C. J. et al. (2015). "The deuterostome context of chordate origins". In: *Nature* 520.7548, pp. 456–465.

Luperchio, T. R., X. Wong, and K. L. Reddy (2014). "Genome regulation at the peripheral zone: Lamina associated domains in development and disease". In: *Current Opinion in Genetics and Development* 25.1, pp. 50–61.

Lupiañez, D. G. et al. (2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions". In: *Cell* 161.5, pp. 1012–1025.

Lupien, M. et al. (2008). "FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription". In: *Cell* 132.6, pp. 958–970.

Lynch, V. J. and G. P. Wagner (2008). "Resurrecting the role of transcription factor change in developmental evolution". In: *Evolution* 62.9, pp. 2131–2154.

Maeso, I. and J. J. Tena (2016). "Favorable genomic environments for cis-regulatory evolution: A novel theoretical framework." eng. In: *Seminars in cell & developmental biology* 57, pp. 2–10.

Maeso, I. et al. (2016). "Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals." In: *BMC biology* 14.1, p. 45.

Mallarino, R. et al. (2011). "Two developmental modules establish 3D beak-shape variation in Darwin's finches." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.10, pp. 4057–62.

Mallo, M. (2018). "Reassessing the Role of Hox Genes during Vertebrate Development and Evolution." In: *Trends in genetics* 34.3, pp. 209–217.

Mansfield, J. H. et al. (2015). "Development of somites and their derivatives in amphioxus, and implications for the evolution of vertebrate somites". In: *EvoDevo* 6.1, pp. 1–30.

Manzanares, M et al. (2000). "Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head." In: *Nature* 408.6814, pp. 854–857.

Marco-Sola, S. et al. (2012). "The GEM mapper: Fast, accurate and versatile alignment by filtration". In: *Nature Methods* 9.12, pp. 1185–1188.

Marinić, M. et al. (2013). "An Integrated Holo-Enhancer Unit Defines Tissue and Gene Specificity of the Fgf8 Regulatory Landscape". In: *Developmental Cell* 24.5, pp. 530–542.

Marlétaz, F. et al. (2018). "Amphioxus functional genomics and the origins of vertebrate gene regulation." In: *Nature* 564.7734, pp. 64–70.

Marti-Renom, M. A. and L. A. Mirny (2011). "Bridging the resolution gap in structural modeling of 3D genome organization". In: *PLoS Computational Biology* 7.7, pp. 1–6.

Martín-Durán, J. M. et al. (2012). "Deuterostomic development in the protostome Priapulus caudatus." In: *Current Biology* 22.22, pp. 2161–6.

Martín-Durán, J. M. et al. (2016). "The developmental basis for the recurrent evolution of deuterostomy and protostomy". In: *Nature Ecology & Evolution* 1.1, p. 0005.

Martín-Durán, J. M. et al. (2018). "Convergent evolution of bilaterian nerve cords". In: *Nature* 553.7686, pp. 45–50.

Martinson, A. S. et al. (2014). "Functional evolution of Erg potassium channel gating reveals an ancient origin for IKr". In: *Proceedings of the National Academy of Sciences* 111.15, pp. 5712–5717.

McGinnis, W. and R. Krumlauf (1992). "Homeobox genes and axial patterning". In: *Cell* 68.2, pp. 283–302.

McLean, C. Y. et al. (2010). "GREAT improves functional interpretation of cis-regulatory regions". In: *Nature Biotechnology* 28.5, pp. 495–501.

Medeiros, D. M. (2013). "The evolution of the neural crest: New perspectives from lamprey and invertebrate neural crest-like cells". In: *Wiley Interdisciplinary Reviews: Developmental Biology* 2.1, pp. 1–15.

Mendel, G. (1946). *Experiments in Plant Hybridization*. Cambridge, MA: Harvard University Press.

Merabet, S. and R. S. Mann (2016). "To Be Specific or Not: The Critical Relationship Between Hox And TALE Proteins". In: *Trends in Genetics* 32.6, pp. 334–347.

Miyamoto, N. and H. Wada (2013). "Hemichordate neurulation and the origin of the neural tube". In: *Nature Communications* 4, pp. 1–8.

Mizuguchi, T. et al. (2014). "Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe." In: *Nature* 516.7531, pp. 432–5.

Montavon, T. et al. (2011). "A regulatory archipelago controls hox genes transcription in digits". In: *Cell* 147.5, pp. 1132–1145.

Morgan, T. (1916). *Sex-linked inheritance in Drosophila*. Washington D.C.: Carnegie Institution of Washington, pp. 120–122.

Mumbach, M. R. et al. (2016). "HiChIP: efficient and sensitive analysis of protein-directed genome architecture." In: *Nature methods* 13.11, pp. 919–922.

Nagano, T. et al. (2017). "Cell-cycle dynamics of chromosomal organization at single-cell resolution". In: *Nature* 547.7661, pp. 61–67.

Nakamura, T. et al. (2016). "Digits and fin rays share common developmental histories". In: *Nature* 537.7619, pp. 225–228.

Nicolás-Pérez, M. et al. (2016). "Analysis of cellular behavior and cytoskeletal dynamics reveal a constriction mechanism driving optic cup morphogenesis". In: *eLife* 5, pp. 1–24.

Noordermeer, D. et al. (2014). "Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci". In: *eLife* 2014.3, pp. 1–21.

Nora, E. P. et al. (2012). "Spatial partitioning of the regulatory landscape of the X-inactivation centre". In: *Nature* 485, pp. 381–385.

Nora, E. P. et al. (2017). "Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization". In: *Cell* 169.5, 930–944.e22.

Ohno, S (1970). *Evolution by gene duplication*. New York: Springer-Verlag.

Onimaru, K. et al. (2011). "Development and evolution of the lateral plate mesoderm: Comparative analysis of amphioxus and lamprey with implications for the acquisition of paired fins". In: *Developmental Biology* 359.1, pp. 124–136.

Onuma, Y. et al. (2002). "Conservation of Pax 6 function and upstream activation by Notch signaling in eye development of frogs and flies". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.4, pp. 2020–2025.

Ou, H. D. et al. (2017). "ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells". In: *Science* 357.6349, eaag0025.

Ovcharenko, I. et al. (2005). "Evolution and functional classification of vertebrate gene deserts". In: *Genome Research* 15.1, pp. 137–145.

Pajni-Underwood, S et al. (2007). "BMP signals control limb bud interdigital programmed cell death by regulating FGF signaling". In: *Development* 134.12, pp. 2359–2368.

Palstra, R.-J. et al. (2003). "The beta-globin nuclear compartment in development and erythroid differentiation." In: *Nature genetics* 35.2, pp. 190–194.

Pani, A. M. et al. (2012). "Ancient deuterostome origins of vertebrate brain signalling centres". In: *Nature* 483.7389, pp. 289–294.

Parker, H. J., M. E. Bronner, and R. Krumlauf (2014). "Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates". In: *Nature* 514.7253, pp. 490–493.

— (2016). "The vertebrate Hox gene regulatory network for hindbrain segmentation: Evolution and diversification: Coupling of a Hox gene regulatory network to hindbrain segmentation is an ancient trait originating at the base of vertebrates H. J. Parker et al." In: *BioEssays* 38.6, pp. 526–538.

Parker, H. J., I. Pushel, and R. Krumlauf (2018). "Coupling the roles of Hox genes to regulatory networks patterning cranial neural crest". In: *Developmental Biology* March, pp. 1–12.

Pascual-Anaya, J. et al. (2012). "Broken colinearity of the amphioxus Hox cluster." In: *EvoDevo* 3.1, p. 28.

Pascual-Anaya, J. et al. (2018). "Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates". In: *Nature Ecology and Evolution* 2.5, pp. 859–866.

Patthey, C., G. Schlosser, and S. M. Shimeld (2014). "The evolutionary history of vertebrate cranial placodes – I: Cell type evolution". In: *Developmental Biology* 389.1, pp. 82–97.

Picco, A. et al. (2017). "The In Vivo Architecture of the Exocyst Provides Structural Basis for Exocytosis". In: *Cell* 168.3, 400–412.e18.

Picelli, S. et al. (2014). "Tn5 transposase and tagmentation procedures for massively scaled sequencing projects." In: *Genome research* 24.12, pp. 2033–2040.

Pinkel, D et al. (1988). "Fluorescence in situ hybridization with human chromosome-specific libraries: Detection of trisomy 21 and translocations of chromosome 4 (tumor cytogenetics/prenatal diagnosis/aneuploidy/interphase)". In: *Proc. Nati. Acad. Sci. USA* 85.December, pp. 9138–9142.

Prud'homme, B. et al. (2006). "Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene". In: *Nature* 440.7087, pp. 1050–1053.

Prykhozhij, S. V., A. Marsico, and S. H. Meijsing (2013). "Zebrafish Expression Ontology of Gene Sets (ZEOGS): A Tool to Analyze Enrichment of Zebrafish Anatomical Terms in Large Gene Sets". In: *Zebrafish* 10.3, pp. 303–315.

Rada-Iglesias, A. et al. (2011). "A unique chromatin signature uncovers early developmental enhancers in humans." In: *Nature* 470.7333, pp. 279–283.

Raff, R. A. (1996). *The Shape of Life: Genes, Development, and the Evolution of Animal Form.* University of Chicago Press.

Rao, S. S. P. et al. (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". In: *Cell* 159.7, pp. 1665–1680.

Rao, S. S. P. et al. (2017). "Cohesin Loss Eliminates All Loop Domains." In: *Cell* 171.2, 305–320.e24.

Ren, R. et al. (2018). "Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms". In: *Molecular Plant* 11.3, pp. 414–428.

Richmond, T. J. et al. (1997). "Crystal structure of the nucleosome core particle at 2.8 {Å}resolution". In: *Nature* 389.6648, pp. 251–260.

Rodríguez-Carballo, E. et al. (2017). "The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes". In: *Genes & Development* 31.22, pp. 2264–2281.

Rokas, A. (2008). "The Origins of Multicellularity and the Early History of the Genetic Toolkit For Animal Development". In: *Annual Review of Genetics* 42.1, pp. 235–251.

Rosenberg, A. B. et al. (2018). "Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding". In: *Science* 360.6385, pp. 176–182.

Roskov, Y. et al. (2018). *Species 2000 & ITIS Catalogue of Life, 2016 Annual Checklist*.

Rowley, M. J. et al. (2017). "Evolutionarily Conserved Principles Predict 3D Chromatin Organization." In: *Molecular cell* 67.5, 837–852.e7.

Russel, D. et al. (2012). "Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies". In: *PLoS Biology* 10.1, e1001244.

Sanborn, A. L. et al. (2015). "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes". In: *Proceedings of the National Academy of Sciences* 112.47, p. 201518552.

Sánchez-Higueras, C. and J. C.-G. Hombría (2016). "Precise long-range migration results from short-range stepwise migration during ring gland organogenesis." In: *Developmental biology* 414.1, pp. 45–57.

Sander, K. (1976). "Specification of the Basic Body Pattern in Insect Embryogenesis". In: *Advances in Insect Physiology* 12.C, pp. 125–238.

Santos, M. E. et al. (2017). "Taxon-restricted genes at the origin of a novel trait allowing access to a new environment". In: *Science* 358.6361, pp. 386–390.

Schlosser, G., C. Patthey, and S. M. Shimeld (2014). "The evolutionary history of vertebrate cranial placodes II: Evolution of ectodermal patterning". In: *Developmental Biology* 389.1, pp. 98–119.

Schones, D. E. et al. (2008). "Dynamic Regulation of Nucleosome Positioning in the Human Genome". In: *Cell* 132.5, pp. 887–898.

Schroeter, E. H., J. A. Kisslinger, and R. Kopan (1998). "Notch-1 signalling requires ligand-induced proteolytic release of intracellular domain". In: *Nature* 393.May, pp. 382–386.

Schuettengruber, B. et al. (2017). "Genome Regulation by Polycomb and Trithorax: 70 Years and Counting". In: *Cell* 171.1, pp. 34–57.

Schwager, E. E. et al. (2017). "The house spider genome reveals an ancient whole-genome duplication during arachnid evolution". In: *BMC Biology* 15.1, pp. 1–27.

Schwaiger, M. et al. (2014). "Evolutionary conservation of the eumetazoan gene regulatory landscape". In: *Genome Research* 24.4, pp. 639–650.

Schwanhüusser, B. et al. (2011). "Global quantification of mammalian gene expression control". In: *Nature* 473.7347, pp. 337–342.

Schwarzer, W. et al. (2017). "Two independent modes of chromatin organization revealed by cohesin removal." In: *Nature* 551.7678, pp. 51–56.

Sebé-Pedrós, A. et al. (2016). "The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity". In: *Cell* 165.5, pp. 1224–1237.

Sebé-Pedrós, A. et al. (2018). "Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq." In: *Cell* 173.6, 1520–1534.e20.

Serra, F. et al. (2017). "Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors". In: *PLoS Computational Biology* 13.7, pp. 1–17.

Sestak, M. S. and T. Domazet-Loso (2015). "Phylostratigraphic profiles in zebrafish uncover chordate origins of the vertebrate brain". In: *Molecular Biology and Evolution* 32.2, pp. 299–312.

Sexton, T. et al. (2012). "Three-dimensional folding and functional organization principles of the Drosophila genome." In: *Cell* 148.3, pp. 458–72.

Sherwood, D. R. and D. R. McClay (1999). "LvNotch signaling mediates secondary mesenchyme specification in the sea urchin embryo." In: *Development* 126.8, pp. 1703–13.

Shiraki, T et al. (2003). "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage". In: *Proceedings of the National Academy of Sciences* 100.26, pp. 15776–15781.

Simakov, O. et al. (2015). "Hemichordate genomes and deuterostome origins". In: *Nature* 527.7579, pp. 459–465.

Slack, J. M., P. W. Holland, and C. F. Graham (1993). "The zootype and the phylotypic stage". In: *Nature* 361.6412, pp. 490–492.

Smemo, S. et al. (2014). "Obesity-associated variants within FTO form long-range functional connections with IRX3." In: *Nature* 507.7492, pp. 371–375.

Smith, J. J. et al. (2018). "The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution". In: *Nature Genetics* 50.2, pp. 270–277.

Spieler, D. et al. (2014). "Restless Legs Syndrome-Associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon". In: *Genome Research* 24.4, pp. 592–603.

Spitz, F. et al. (2001). "Large scale transgenic and cluster deletion analysis of the HoxD complex separate an ancestral regulatory module from evolutionary innovations". In: *Genes and Development* 15.17, pp. 2209–2214.

Stemple, D. L. (2005). "Structure and function of the notochord: an essential organ for chordate development". In: *Development* 132.11, pp. 2503–2512.

Stolfi, A. et al. (2010). "Early Chordate Origins of the Vertebrate Second Heart Field". In: *Science* 329.5991, pp. 565–568.

Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." In: *Nature* 403.6765, pp. 41–45.

Symmons, O. et al. (2016). "The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances." In: *Developmental cell* 39.5, pp. 529–543.

Takaku, M. et al. (2016). "GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler". In: *Genome Biology* 17.1, pp. 1–16.

Tanaka, M. (2016). "Fins into limbs: Autopod acquisition and anterior elements reduction by modifying gene networks involving 5'Hox, Gli3, and Shh". In: *Developmental Biology* 413.1, pp. 1–7.

Tarazona, O. A. et al. (2018). "Evolution of limb development in cephalopod mollusks". In: *bioRxiv*.

Tena, J. J. et al. (2011). "An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation." In: *Nature communications* 2.1, p. 310.

Tena, J. J. et al. (2014). "Comparative epigenomics in distantly related teleost species identifies conserved cis-regulatory nodes active during the vertebrate phylotypic period". In: *Genome Research* 24.7, pp. 1075–1085.

Tie, F. et al. (2016). "Polycomb inhibits histone acetylation by CBP by binding directly to its catalytic domain". In: *Proceedings of the National Academy of Sciences* 113.6, E744–E753.

Tolhuis, B. et al. (2002). "Looping and interaction between hypersensitive sites in the active beta-globin locus". In: *Molecular Cell* 10.6, pp. 1453–1465.

Tomoyasu, Y., S. R. Wheeler, and R. E. Denell (2005). "Ultrabithorax is required for membranous wing identity in the beetle Tribolium castaneum." In: *Nature* 433.7026, pp. 643–7.

Tschopp, P., A. J. Christen, and D. Duboule (2012). "Bimodal control of Hoxd gene transcription in the spinal cord defines two regulatory subclusters". In: *Development* 139.5, pp. 929–939.

Tschopp, P. and C. J. Tabin (2017). "Deep homology in the age of next-generation sequencing". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1713.

Venkatesh, B. et al. (2014). "Elephant shark genome provides unique insights into gnathostome evolution". In: *Nature* 505.7482, pp. 174–179.

Vicente-García, C. et al. (2017). "Regulatory landscape fusion in rhabdomyosarcoma through interactions between the PAX3 promoter and FOXO1 regulatory elements". In: *Genome Biology* 18.1, p. 106.

Vietri Rudan, M. et al. (2015). "Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture". In: *Cell Reports* 10.8, pp. 1297–1309.

Vieux-Rochas, M. et al. (2015). "Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain". In: *Proceedings of the National Academy of Sciences* 112.15, pp. 4672–4677.

Visel, A, E. M. Rubin, and L. A. Pennacchio (2009). "Genomic views of distant-acting enhancers". In: *Nature* 461.7261, pp. 199–205.

Visel, A. et al. (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers". In: *Nature* 457.7231, pp. 854–858.

Voigt, P., W. W. Tee, and D. Reinberg (2013). "A double take on bivalent promoters". In: *Genes and Development* 27.12, pp. 1318–1338.

Von Ohlen, T et al. (1997). "Hedgehog signaling regulates transcription through cubitus interruptus, a sequence-specific DNA binding protein." In: *Proceedings of the National Academy of Sciences of the United States of America* 94.6, pp. 2404–2409.

Wagner, G. P. (2007). "The developmental genetics of homology." In: *Nature reviews. Genetics* 8.6, pp. 473–9.

Wan, L.-B. et al. (2008). "Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development". In: *Development* 135.16, pp. 2729–2738.

Wang, C. et al. (2015). "Genome-wide analysis of local chromatin packing in Arabidopsis thaliana". In: *Genome Research* 25.2, pp. 246–256.

Wang, X. et al. (2014). "Genome-wide and single-base resolution DNA methylomes of the Pacific oyster Crassostrea gigas provide insight into the evolution of invertebrate CpG methylation." In: *BMC genomics* 15.1, p. 1119.

Wang, Z., M. Gerstein, and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." In: *Nature reviews. Genetics* 10.1, pp. 57–63.

Wang, Z. et al. (2013). "The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan." In: *Nature genetics* 45.6, pp. 701–706.

Watson, J. D. (1963). "Involvement of RNA in the synthesis of proteins". In: *Science* 140.3562, pp. 17–26.

Watson, J. D. and F. H. C. Crick (1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171, p. 737.

Weatherbee, S. D. et al. (1998). "Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the Drosophila haltere". In: *Genes & Development* 12.10, pp. 1474–1482.

Weintraub, A. S. et al. (2017). "YY1 Is a Structural Regulator of Enhancer-Promoter Loops." In: *Cell* 171.7, 1573–1588.e28.

Werken, H. J. G. van de et al. (2012). "Robust 4C-seq data analysis to screen for regulatory DNA interactions". In: *Nature Methods* 9.10, pp. 969–972.

Wijesena, N., D. K. Simmons, and M. Q. Martindale (2017). "Antagonistic BMP-cWNT signaling in the cnidarian Nematostella vectensis reveals insight into the evolution of mesoderm." In: *Proceedings of the National Academy of Sciences of the United States of America* 114.28, E5608–E5615.

Williams, T. A. (1994). "The nauplius larva of crustaceans: Functional diversity and the phylotypic stage". In: *Integrative and Comparative Biology* 34.4, pp. 562–569.

Willmore, K. E. (2012). "The Body Plan Concept and Its Centrality in Evo-Devo". In: *Evolution: Education and Outreach* 5.2, pp. 219–230.

Woltering, J. M. et al. (2014). "Conservation and divergence of regulatory strategies at Hox Loci and the origin of tetrapod digits." In: *PLoS biology* 12.1, e1001773.

Xue, S. et al. (2015). "RNA regulons in Hox 5' UTRs confer ribosome specificity to gene regulation". In: *Nature* 517.7532, pp. 33–38.

Ye, T. et al. (2011). "seqMINER: an integrated ChIP-seq data interpretation platform." In: *Nucleic acids research* 39.6, e35.

Zabidi, M. a. et al. (2014). "Enhancer—core-promoter specificity separates developmental and housekeeping gene regulation". In: *Nature* 518.7540, pp. 556–559.

Zhang, Y. et al. (2008). "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9, R137.

Zullo, J. M. et al. (2012). "DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina". In: *Cell* 149.7, pp. 1474–1487.

Index of abbreviations

3C Chromosome Conformation Capture. Experiment used to assess how much do two loci interact in space with each other in the nucleus.

4C-seq Circular Chromosome Conformation Capture. Experiment used to assess how much a given locus called bait interact with the rest of the loci of the genome.

5C Carbon-Copy Chromosome Conformation Capture. Experiment used to assess the many loci vs many loci interactions in an specific genomic area.

AER Apical Ectodermal Ridge. Signalling center of both limb and fin buds, located at their most distal parts. It is a source of Fgf and governs the distal growth of the structure.

AP axis Antero-posterior axis.

ATAC-seq Assay for Transposase Accessible Chromatin. Experiment that uses a hyperactive transposase in native chromatin in order to predict open chromatin regions and therefore CREs like promoters and enhancers.

CAGE-seq Cap Analysis of Gene Expression. Experiment that reveal the whole set of 5' ends of the mRNAs of a transcriptome, allowing to establish the exact position of the different TSSs.

Capture-C Technique derived from HiC that uses oligonucleotide probes in order to obtain a high resolution HiC contact map in specific regions.

CDS Coding sequence. The part of a gene that contains the codons that encode the aminoacidic sequence of proteins.

ChIA-PET . Technique derived from HiC that allows to identify the 3D interactions that are mediated by a given protein of interest using antibodies.

ChIN Character identity network. Modular structure inside a developmental GRN that ensures that the fate of a given cell population is going to be to become a specific character, for instance an eye.

CHIP-seq Chromatin Immunoprecipitation Sequencing. Experiment that allows to identify where in the genome an specific protein recognizable by an antibody is found.

chromEMT Chromosome Electron Microscopy Tomography. Electron microscopy technique that allows to look at the folding of genomes at different resolutions.

CID Chromatin Interacting Domains. Regions inside bacterial chromosomes containing loci that tend to interact more among themselves than with loci outside the CID. See TADs.

CNS Central Nervous System.

CPF Cardiopharyngeal field. Cell population found in vertebrates that give rise both to specific muscles of the heart and the head.

CRE Cis-regulatory element. Non-coding sequence of the genome that participate in the transcriptional regulation of genes. Promoters and enhancers, for instance, are CREs.

DPE Downstream Promoter Element. It is a *Drosophila* specific promoter sequence located downstream from the TSS.

ES cells/ESC Embryonic Stem Cells.

FISH Fluorescent In-Situ Hybridization. Technique that, among other applications, allows to identify specific genomic loci in light microscopy preparations.

GRN Gene Regulatory Network. Theoretical model that allows to understand how different genes establish regulatory interactions among them.

GRO-seq Global Run-On Sequencing. Technique that allows to identify active spots of transcription genome wide.

H3K27ac Acetylation of the lysine in position 27 of the histone 3. Epigenetic modification of the nucleosomes, typically found around active promoters and enhancers.

H3K27me3 Trimethylation of the lysine in position 27 of the histone 3. Epigenetic modification of the nucleosomes typically found around promoters represed by the PRC complex.

H3K4me1 Monomethylation of the lysine in position 4 of the histone 3. Epigenetic modification typically found around both poised and active enhancers.

H3K4me3 Monomethylation of the lysine in position 4 of the histone 3. Epigenetic modification typically found around active promoters.

H3K9me2/H3K9me3 Di or trimethylation of the lysine in position 9 of the histone 3. Epigenetic modification typically found in heterochromatin silenced regions.

HiC C-technique that allows to identify the 3D interactions of every pair of loci in the genome.

HiChIP Technique derived from HiC conceptually similar to ChIA-PET but with a higher yield of contacts.

LAD Lamin Associated Domains. Region of the chromosome that is associated to the nuclear lamina and therefore is typically transcriptionally silent.

LCA Last Common Ancestor.

LCR Locus Control Region. Distal enhancer that regulate several of the β-globin promoters by looping.

MethylC-seq Technique that uses NGS in order to assess the levels of CpG methylation genome wide.

NGS Next Generation Sequencing.

PC Principal Component.

PIC Pre-Initiation Complex. Complex formed both by general TFs and the RNAP II at the onset of transcription.

PRC Polycomb Repressive Complex. Group of proteins that are recruited to specific loci in order to repress transcription. Some of them are responsible for the H3K27me3 epigenetic modification.

PRE Polycomb Response Elements. *Drosophila* specific sequences that are able to recruit the PRC.

PSM Presomitic mesoderm. Cell population that give rise to the somites of vertebrate embryos.

RA Retinoic acid. Important developmental morphogen.

RL Regulatory Landscape. Genomic region that is accessible in 3D by a given promoter.

RNAP II RNA Polymerase II. Polymerase responsible for the transcription of mRNAs.

STARR-seq Self-Transcribing Active Regulatory Region Sequencing. Experiment that assays the enhancer activity of a genome wide library of DNA inserts in a specific cellular context.

T2/T3 Thoracic segments 2 and 3 of insects.

TAD Topologically Associated Domains. 3D isolated genomic regions in which most animal genomes are compartmentalized. Enhancer-promoter contacts are favored between loci belonging to the same TAD.

TF Transcription Factor. Protein that binds to CREs of the DNA in order to active or repress the transcription of a given gene.

TSS Transcriptional Start Site. Precise base pair where the transcription of a RNA molecule begins.

UTR Untranslated region. Part of the mRNA sequence that does not contain codons encoding aminoacids. They regulate the binding of the RNAs to the ribosomes and also the stability of the molecule.

WGD Whole Genome Duplication. Mutational event that results in the complete duplication of the whole genome (presumably due to meiotic problems).

ZPA Zone of Polarizing Activity. Signalling center located in the posterior end of both limb and fin buds. It is a source of the morphogen Shh that helps to establish the antero-posterior polarity of the appendages.

ZRS Zone of Polarizing Activity Regulatory Sequence. *Shh* enhancer that drives the expression of this gene in the ZPA.

List of Figures

List of Tables